If CLIP Could Talk: Understanding Vision-Language Model Representations Through Their Preferred Concept Descriptions

Reza Esfandiarpoor[1] **Cristina Menghini**[2] **Stephen H. Bach**[1]
[1]Department of Computer Science, Brown University
[2]Data Science Institute, Brown University
{reza_esfandiarpoor,cristina_menghini,stephen_bach}@brown.edu

Abstract

Recent works often assume that Vision-Language Model (VLM) representations are based on visual attributes like shape. However, it is unclear to what extent VLMs prioritize this information to represent concepts. We propose Extract and Explore (EX2), a novel approach to characterize textual features that are important for VLMs. EX2 uses reinforcement learning to align a large language model with VLM preferences and generates descriptions that incorporate features that are important for the VLM. Then, we inspect the descriptions to identify features that contribute to VLM representations. Using EX2, we find that spurious descriptions have a major role in VLM representations despite providing no helpful information, e.g., `Click to enlarge photo of CONCEPT`. More importantly, among informative descriptions, VLMs rely significantly on non-visual attributes like habitat (e.g., North America) to represent visual concepts. Also, our analysis reveals that different VLMs prioritize different attributes in their representations. Overall, we show that VLMs do not simply match images to scene descriptions and that non-visual or even spurious descriptions significantly influence their representations. Code: `https://github.com/BatsResearch/ex2`

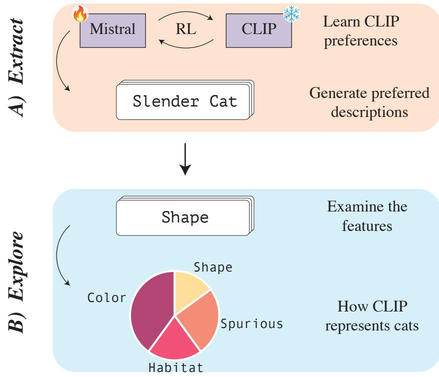


Figure 1: Extract: we align Mistral with VLM preferences and generate descriptions that contain features that are important for the VLM. Explore: we examine various aspects of these descriptions to identify features that contribute to VLM representations.

1 Introduction

The ability of contrastive Vision-Language Models (VLMs) to match related text and images depends significantly on the content of text descriptions (Novack et al., 2023; Radford et al., 2021). Heuristically, to better identify the related images, humans seek more detailed information about the physical appearance of concepts. Many recent works extend this heuristic to VLMs and augment the descriptions (also called prompts) with additional visual information (e.g., shape) to more accurately identify the related images (Menon and Vondrick, 2023; Yan et al., 2023). However, in practice, even augmenting the descriptions with random words improves the performance (Roth et al., 2023), and it is unclear what factors contribute to VLM concept representations. Here, we propose a novel analysis method for characterizing textual features that contribute to the VLM representation of a concept. Simply put, our goal is to understand how VLMs encode different concepts.

Despite the growing body of work on various aspects of VLMs (e.g., sensitivity to word order) (Akula et al., 2020; Ma et al., 2023; Yuksekgonul et al., 2022), there is limited understanding of VLMs' world knowledge acquired during pretraining. Yun et al. (2021) show that VLMs and Large Language Models (LLMs) perform similarly on language tasks that require knowledge of the physical world. *If VLMs do not learn about the physical world from images, what do they learn?* Moreover, Yun et al. (2022) show that, the VLM representation of a concept is often not based on its visual attributes like color and shape. *If not based on visual attributes, how do VLMs represent concepts?* Our work introduces a new analysis method for studying such exploratory questions.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 9797–9819
November 12-16, 2024 ©2024 Association for Computational Linguistics

Unlike previous works that focus on a very specific question, our goal is to explore and characterize the features that contribute to VLM representations with the feedback from the VLM itself.

In this paper, we propose Extract and Explore (EX2), a novel analysis method for understanding how VLMs represent concepts. Instead of directly probing the VLM, EX2 uses reinforcement learning (RL) to align an LLM with VLM preferences. Our reward function measures the similarity between the concept description and its images. Thus, it encourages the LLM to generate descriptions that incorporate features that are important for the VLM. Then, we can inspect these descriptions from various perspectives to identify common factors that contribute to VLM representations. In this work, we examine the descriptions to understand how often VLMs prioritize additional information to represent concepts and how often this information explains the physical appearance of concepts.

We use EX2 to analyze seven different VLMs on six fine-grained classification datasets. The aligned LLM successfully learns features that are important for the VLM and generates descriptions that achieve better classification accuracy than a set of generic descriptions. We find that spurious descriptions (e.g., `Click to enlarge photo of CONCEPT`) have a major role in VLM representations despite providing no helpful information about concepts. More importantly, even when VLMs prioritize informative descriptions, they significantly rely on non-visual attributes[1] like habitat (e.g., North America) to represent visual concepts. In our fine-grained analysis, we find that different VLMs represent similar concepts differently, i.e., prioritize different attributes. Even the same VLM prioritizes different attributes for different datasets, suggesting that extensive experiments across different datasets are required to draw conclusions about VLMs. EX2 is specifically suited to automatically carry out such analysis on a wide range of existing classification datasets. Finally, we show EX2's application for hypothesis generation and suggest new research questions based on our findings.

Our findings encourage further work on image-text pre-training to address VLMs' reliance on spurious descriptions and the alignment between non-visual information and images. Moreover, EX2 provides the future work with a flexible tool to study the impact of different pre-training methods and datasets from various aspects. We summarize our contributions and findings as following:

- We propose EX2, a novel analysis method that aligns an LLM with VLM preferences in order to characterize textual features that contribute to VLM representations.
- We show that EX2 successfully learns features that are important for the VLM and generates descriptions that improve the downstream classification accuracy, highlighting the benefits of adapting to VLM preferences for downstream tasks as well as analysis.
- We show that VLMs significantly rely on spurious or non-visual descriptions to represent visual concepts. We find that different VLMs prioritize different attributes to represent similar concepts. Even the same VLM prioritizes different attributes across datasets, emphasizing the benefits of EX2's ability to automatically analyze VLMs on a wide range of existing classification datasets.

2 Related Work

Reinforcement Learning for Language Models In recent years, RL has been successfully used to align LLMs with human preferences, known as reinforcement learning with human feedback (RLHF) (Bai et al., 2022; Ramamurthy et al., 2023). Instead of human preferences, we use similar methods to Stiennon et al. (2020) and Ziegler et al. (2019) to align the LLM with VLM's preferences for descriptions that it deems more accurate. Unlike previous methods, we use preference learning not as our goal but as a tool to understand the knowledge that VLMs acquire during pre-training.

Vision-Language Models (VLMs) We primarily focus on contrastive VLMs, which are trained to push related image-text pairs closer and unrelated pairs farther in the embedding space (Jia et al., 2021; Radford et al., 2021). Multimodal language models (MLMs) are another category of VLMs that use additional training to incorporate the image features into LLMs in order to condition the text generation on both input text and images (Li et al., 2023a; Liu et al., 2023). Despite the many advantages of MLMs, contrastive VLMs are still better suited for tasks like image classification (Alayrac et al., 2022)

[1]Throughout the paper, "attribute" refers to concept attributes like color and shape, while "feature" and "characteristic" refer to properties of descriptions like the type of information they contain or text style.

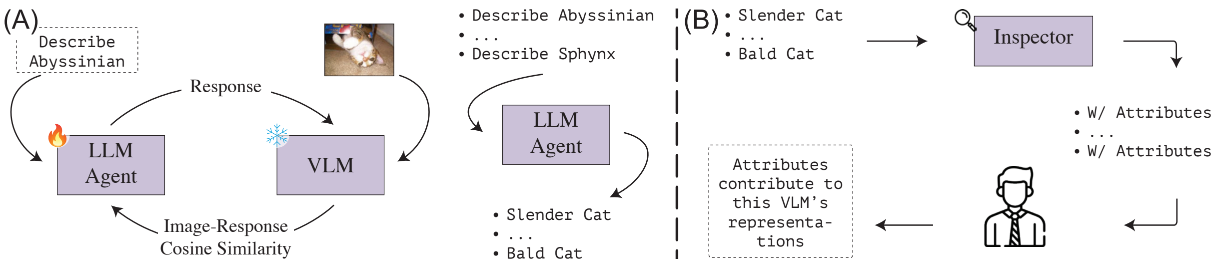


Figure 2: Extract and Explore (EX2) overview. A) We use RL to fine-tune an LLM to generate concept descriptions that are closer to the corresponding images in the VLM embedding space, thus, the descriptions incorporate features that the VLM uses to represent the concepts. We use the aligned LLM to generate the VLM's preferred description for all concepts. B) We inspect these descriptions from various aspects, e.g., if they are informative or describe visual attributes. Based on the aggregate results, we draw conclusions about how the VLM represents concepts.

or creating large search indices for efficient filtering and retrieval (Schuhmann et al., 2022). Contrastive VLMs also serve as the backbone for many other methods or downstream applications like image manipulation (Patashnik et al., 2021). Even most MLMs use contrastive VLMs to extract the image features (Liu et al., 2023). Thus, understanding how contrastive VLMs represent concepts remains an important issue that also impacts other types of VLMs and many downstream applications. We aim to investigate this question without fine-tuning or modifying the VLM after pre-training.

Detailed Descriptions for VLMs One line of work provides VLMs with detailed concept descriptions to improve the classification accuracy (Feng et al., 2023; Li et al., 2023b; Pratt et al., 2023; Yan et al., 2023; Yang et al., 2023). For instance, Menon and Vondrick (2023) suggest augmenting the descriptions with concept attributes. Esfandiarpoor and Bach (2024) further improve this idea and ensure the attributes differentiate the target classes. However, Roth et al. (2023) question the role of additional information and suggest that even adding random words and characters to descriptions leads to similar improvements. Here, we propose a novel model analysis method to study the role of additional information in VLM representations.

Vision-Language Model Analysis The massive success of VLMs has sparked an interest in understanding how they interpret text descriptions. Previous works have studied VLMs from various aspects such as the relative importance of verbs and nouns (Hendricks and Nematzadeh, 2021), VLMs' sensitivity to word order (Akula et al., 2020), linguistic features of descriptions (Castro et al., 2023), and, the most popular, their poor compositional capabilities (Hsieh et al., 2023; Lewis et al., 2022; Ma et al., 2023; Parcalabescu et al., 2022; Schiappa et al., 2023; Thrush et al., 2022; Xu et al., 2024). Although this line of work provides important information about specific aspects of VLMs' behavior, our understanding of the world knowledge that VLMs acquire during pre-training remains limited. As discussed in Section 1, this becomes even more important considering that previous works suggest VLMs' behavior deviates from the expected results of vision and language pre-training (Yun et al., 2022, 2021). Given the importance of concept representations (Lovering and Pavlick, 2022; Merullo et al., 2022; Patel and Pavlick, 2021; Pavlick, 2022), our goal is to understand what VLMs learn during pre-training and characterize factors contributing to their concept representations.

Moreover, we propose a new paradigm for studying VLMs that allows us to conduct such exploratory analysis. Currently, the most common approach for VLM analysis relies on custom datasets that test a specific hypothesis. However, our approach is compatible with existing classification datasets and is not tied to a specific hypothesis. We can even use it to generate new hypotheses (Section 4.6). There is also a complementary line of work for understanding VLM representations in terms of images instead of text (Ghiasi et al., 2022; Kazemi et al., 2024). For natural language tasks, Perez et al. (2022) use a similar method, i.e., LLM alignment, to discover the harmful generations of other LLMs. We emphasize that our method is not meant to replace existing VLM analysis approaches but to provide a complementary tool that allows for exploratory model analysis.

3 Extract and Explore

Given the large number of textual features that could contribute to VLM representations (e.g., concept attributes and linguistic patterns), it is very costly, if not impossible, to curate and probe VLMs for an exhaustive set of features. Instead, EX2 learns to automatically generate concept descriptions that incorporate textual features that are important for the VLM, i.e., Extract VLM's preferred descriptions (Fig. 2a). Then, EX2 inspects these descriptions for common patterns to identify features that the VLM uses to represent concepts, i.e., Explore VLM's preferred descriptions (Fig. 2b).

3.1 Extracting VLM's Preferred Descriptions

A Large Search Space To expand the search space for potential features, we use LLMs to generate concept descriptions. LLM's ability to control various aspects of text generation, such as world knowledge and text style (Jiang et al., 2020; Petroni et al., 2019), leads to a more thorough search than a predefined set of features. To further expand the search space, we use 25 different questions to query the LLM about various aspects of each concept (refer to Appendix C for the list of question templates).

LLM Alignment with VLM Preferences To increase the likelihood of VLM's desired features in descriptions, we use reinforcement learning to align the LLM with VLM preferences. Thus, we define a reward function that gives a higher score to descriptions that include features that are important for the VLM. Intuitively, from the VLM's perspective, the reward function encourages more accurate concept descriptions. Since contrastive VLMs like CLIP are trained to push related text and images closer, VLMs deem a concept description accurate if it is close to the embedding of the corresponding images. Therefore, we define the reward function as the average cosine similarity between the concept description and concept images.

Specifically, for some description d_c and a set of images D_c for class c, we calculate the reward as:

$$R(d_c) = \frac{1}{|D_c|} \sum_{x \in D_c} \tau \cos(\Phi_I(x), \Phi_T(d_c)) - \beta KL \,,$$

where Φ_I and Φ_T are the VLM image and text encoders, respectively, τ is a constant scaling factor, and β is the coefficient for the KL divergence between the original and current model. The KL penalty encourages the generation of meaningful descriptions that we can later analyze and helps with convergence by constraining the search space. Since we want to compare the generations across experiments, we adopt the adaptive KL coefficient technique of Ziegler et al. (2019) to achieve roughly similar KL divergences across experiments.

With the 25 questions for all classes as the LLM input and this reward function, we use the same method as Stiennon et al. (2020) and Ziegler et al. (2019), which uses Proximal Policy Optimization (Schulman et al., 2017), to update the LLM to increase the reward score.

Then, for each concept, EX2 uses the aligned LLM to generate a set of descriptions that are similar to the corresponding images, i.e., they are the VLM's preferred descriptions for the concept.

3.2 Exploring VLM's Preferred Descriptions

Since the aligned descriptions incorporate features that are important for the VLM, we inspect these descriptions for desired characteristics to understand the extent of their contribution to VLM representations. For instance, we can check if VLMs rely on attributes to represent concepts or if VLMs are biased towards a specific text style. We use ChatGPT (Brown et al., 2020; Ouyang et al., 2022) to automatically inspect the large number of descriptions.

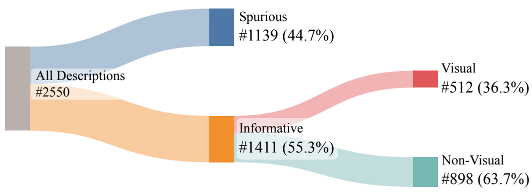


Figure 3: Breakdown of aligned descriptions for CLIP on Flowers. CLIP significantly relies on spurious or non-visual information to represent flower species.

Here, we study VLMs as visual perception tools that are expected to match the description of a concept's physical appearance to images of instances of that concept. Similar to the reward function, our analysis focuses on the physical appearance of concepts rather than the entire content of individual images. For many applications, it is undesirable for models to rely on *spurious* or *non-visual* information. Also see Appendices A.1 and A.2 for extended discussion. We consider descriptions that provide no world knowledge about the concept to be spurious (e.g., `click to enlarge photo of the cat`). We consider informative descriptions to be non-visual if they only contain additional information that is not about the concept's physical appearance (e.g., `The bobolink is a North American bird`). For example, we find that CLIP relies significantly on both spurious and non-visual information to represent flowers, with 45% of descriptions providing no helpful informa-

VLM	Flowers		Pets		CUB		Stanford Dogs		Aircrafts		Stanford Cars	
	Temp.	EX2	Temp.	EX2	Temp.	EX2	Temp.	EX2	Temp.	EX2	Temp.	EX2
CLIP	63.31	**73.65**	84.68	**88.34**	51.48	**53.54**	59.59	**61.49**	21.36	**22.89**	60.15	**60.88**
ALIGN	59.44	**62.19**	82.56	**86.86**	36.81	**39.66**	52.38	**56.24**	12.42	**17.04**	72.73	**74.15**
EVA	74.47	**79.05**	91.80	**94.52**	**74.63**	73.52	76.08	**79.37**	37.08	**39.66**	**91.62**	91.22
SigLIP	82.48	**86.81**	92.10	**94.49**	61.96	**65.33**	76.76	**78.89**	48.21	**50.56**	90.92	**91.22**
MetaCLIP	71.44	**73.18**	87.84	**91.42**	69.88	**69.99**	65.50	**68.87**	**37.50**	36.72	**76.22**	76.17
CLIPA	77.13	**79.61**	91.93	**93.87**	74.56	**76.99**	77.19	**79.34**	39.75	**41.31**	**94.80**	**94.76**
DFN	**88.96**	88.86	94.33	94.33	87.25	**87.78**	84.31	**85.75**	**78.76**	75.88	**96.14**	95.78

Table 1: Accuracy of EX2 descriptions. Temp. is the set of 80 generic templates (Radford et al., 2021).

tion and 64% of informative descriptions explaining non-visual attributes (Fig. 3).

The space of potential hypotheses goes beyond our specific analysis, and we encourage future work to study other aspects of the aligned descriptions, such as text style or linguistic patterns. Furthermore, as we show in Section 4.6, the trends in EX2 descriptions provide helpful cues for discovering new research questions for further investigation.

4 Experiments

In this section, we use EX2 to analyze different VLMs at three levels (Fig. 3). We first validate that the LLM learns features that are important for each VLM, and aligned descriptions improve the classification accuracy. Second, we show that spurious descriptions contribute significantly to VLM representations despite providing no helpful information. More importantly, at the third level, we find that VLMs rely significantly on non-visual attributes to represent visual concepts. We discuss potential correlations between pre-training details and our findings in Appendix A.3. We conduct a more fine-grained analysis that reveals even for the same dataset, different VLMs prioritize different attributes to represent the concepts. Moreover, even the same VLM prioritizes different attributes for different datasets. Finally, we show EX2's application for hypothesis generation and suggest new research questions based on our observations.

4.1 Setup

After RL fine-tuning, we query the LLM with the same questions used during fine-tuning (Section 3) and generate 25 descriptions for each concept that incorporate features that are important for the VLM. To study VLM preferences, our analysis is only based on a subset of generated descriptions that help VLMs better identify the related images and hence are representative of VLM preferences. We consider a set of descriptions helpful if they achieve better image classification accuracy on the corresponding task than the ensemble of 80 generic descriptions used by Radford et al. (2021). Similar to Radford et al. (2021), we use the 25 generated descriptions for each concept as a prompt ensemble and use cosine similarity to predict the label for each image.

Datasets We use six classification datasets for analysis. CUB200-2011: fine-grained bird species recognition (Wah et al., 2011). FGVCAircraft: aircraft model classification (Maji et al., 2013). Flowers102: fine-grained flower species recognition (Nilsback and Zisserman, 2008). Oxford IIIT Pets (Parkhi et al., 2012). Stanford Dogs (Khosla et al., 2011). Stanford Cars (Krause et al., 2013).

Models We use Mistral-7B as our LLM (Jiang et al., 2023). We choose seven different VLMs for our analysis. CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), EVA (Fang et al., 2023b), SigLIP (Zhai et al., 2023), MetaCLIP (Xu et al., 2023), CLIPA (Li et al., 2023c), and DFN (Fang et al., 2023a). Refer to Appendix D for details.

4.2 Successful Alignment

We use classification accuracy to verify that the LLM learns what features contribute to VLM representations. In 33 out of 42 experiments, the LLM successfully learns what features help the VLM to identify the related images, and the aligned descriptions improve the accuracy compared to the generic template set (Table 1). Notably, most of the other nine experiments involve the Stanford Cars dataset or the DFN model. Although our goal is model analysis, our results also emphasize the benefits of adapting to VLM preferences for downstream tasks

	Flowers	Pets	CUB	Stanford Dogs	Aircrafts	Stanford Cars
CLIP	**55.33**	**61.73**	**83.50**	**49.03**	**57.84**	**44.47**
ALIGN	**63.92**	**68.76**	**63.66**	**73.87**	**65.60**	**42.88**
EVA	1.80	0.00	-	**36.60**	**29.20**	-
SigLIP	**44.59**	18.70	23.06	21.00	5.36	9.96
MetaCLIP	**43.06**	**32.22**	**26.58**	**51.40**	-	-
CLIPA	**55.18**	**26.70**	**47.84**	**78.80**	6.24	-
DFN	-	-	0.32	22.90	-	-

Table 2: The percentage of informative descriptions for experiments that the LLM successfully learns the VLM preferences and improves the classification accuracy. Bold numbers are $> 25\%$.

like classification. In the remainder of this section, we analyze the descriptions for the 33 experiments where we successfully learn the VLM preferences to characterize textual features that contribute to concept representations.

4.3 Informative vs Spurious Descriptions

Heuristically, we expect descriptions to better represent concepts when they contain additional information and make little to no difference otherwise. To study the role of spurious descriptions in VLM representations, we inspect if descriptions provide additional information about concepts.

Spurious descriptions are a major factor in VLM representations. Table 2 reports the percentage of descriptions that provide additional information about concepts for the 33 cases in which the LLM successfully learns the VLM preferences. We refer to descriptions that provide no additional information about concepts as spurious (e.g., `Photo of CONCEPT attracted my attention`). We observe that in 10 cases, the improvements are almost solely driven by spurious descriptions, and there is no considerable amount of additional information in descriptions, i.e., <25% of descriptions are informative. In a few of these cases, the LLM learns to only generate the concept names with some artifacts, without any helpful information (see Appendix B for examples). Notably, SigLIP benefits more from spurious descriptions (5/6 datasets) than other VLMs. To understand the role of spurious descriptions in the remaining 23 cases, we separate the informative and non-informative descriptions and measure the classification accuracy (Table 3). Fortunately, VLMs do not solely rely on spurious descriptions, and in 19 cases, the informative descriptions alone improve the performance. However, in 16 cases, spurious descriptions alone lead to a considerable boost in accuracy. In total, in 26 out of the 33 experiments, spurious descriptions are either almost solely responsible for the improvements or have a considerable contribution.

Dataset	VLM	Temp. Set	w/o Info	w/ Info
Flowers	CLIP	63.31	72.19	**73.77**
	ALIGN	59.44	59.39	**61.90**
	SigLIP	82.48	85.43	**87.05**
	MetaCLIP	71.44	**73.44**	70.39
	CLIPA	77.13	79.36	**79.40**
Pets	CLIP	84.68	87.49	**87.90**
	ALIGN	82.56	85.61	**86.54**
	MetaCLIP	87.84	**91.61**	90.22
	CLIPA	91.93	**94.09**	92.75
CUB	CLIP	51.48	46.50	**52.88**
	ALIGN	36.81	36.99	**39.47**
	MetaCLIP	**69.88**	69.43	67.88
	CLIPA	74.56	75.94	**76.73**
Stanford Dogs	CLIP	59.59	**60.71**	60.34
	ALIGN	52.38	52.32	**55.85**
	EVA	76.08	**79.41**	78.07
	MetaCLIP	65.50	**68.54**	66.43
	CLIPA	77.19	76.82	**79.09**
Aircrafts	CLIP	21.36	**23.58**	22.17
	ALIGN	12.42	14.79	**18.09**
	EVA	37.08	**39.09**	36.63
Stanford Cars	CLIP	60.15	**60.17**	59.74
	ALIGN	72.73	73.54	**74.00**

Table 3: Classification accuracy of informative (w/ Info) and spurious (w/o Info) descriptions. First and second best numbers are in bold and underline.

4.4 Visual vs Non-visual Information

VLMs are often thought to match images to scene descriptions. Therefore, it is natural to think of descriptions of visual attributes, like color, as the basis for VLM representations and ignore or assume a negligible role for non-visual attributes like habitat (e.g., North America). To investigate the contribution of non-visual attributes to VLM representations, we inspect if each informative description provides visual or non-visual information.

VLM	Flowers	Pets	Stanford Dogs
CLIP	**36.29**	**54.82**	**33.45**
ALIGN	**56.44**	**55.66**	**55.55**
EVA	-	-	**30.78**
SigLIP	**67.46**	-	-
MetaCLIP	-	**57.38**	**45.40**
CLIPA	**58.71**	**64.78**	**60.03**
	CUB	Aircrafts	Stanford Cars
CLIP	**64.00**	10.24	-
ALIGN	18.69	9.82	8.66
CLIPA	**39.34**	-	-

Table 4: Percentage of informative descriptions that contain visual attributes. Bold numbers are > 25%.

Dataset	VLM	Temp. Set	Vis	Non-Vis
Flowers	CLIP	63.31	66.92	**73.54**
	ALIGN	59.44	60.61	**62.34**
	SigLip	82.48	**86.84**	85.10
	CLIPA	77.13	**79.79**	79.22
Pets	CLIP	84.68	86.18	**87.35**
	ALIGN	82.56	**86.02**	85.69
	MetaCLIP	87.84	**89.51**	86.94
	CLIPA	**91.93**	91.06	91.20
CUB	CLIP	51.48	**53.40**	50.36
	CLIPA	74.56	73.33	**76.42**
Stanford Dogs	CLIP	59.59	55.79	**60.56**
	ALIGN	52.38	**55.12**	53.57
	EVA	76.08	74.37	**78.66**
	MetaCLIP	**65.50**	62.73	65.20
	CLIPA	77.19	**79.07**	78.25

Table 5: Classification accuracy of descriptions that explain visual/non-visual characteristics of the concepts. First and second best numbers are in bold and underline.

Non-visual descriptions contribute significantly to VLM representations. For the 19 cases in which informative descriptions alone lead to a considerable improvement, we report the percentage of informative descriptions that contain visual attributes in Table 4. In four cases, non-visual descriptions are almost solely responsible for the accuracy boosts, and only a small fraction (<25%) of informative descriptions contain visual information. To study the role of non-visual information in the remaining cases, we split the informative descriptions into visual and non-visual categories and measure the classification accuracy (Table 5). Although visual information contributes to VLM representations, in 11 out of 15 cases, non-visual information alone considerably boosts accuracy. In total, in 15 out of 19 cases, non-visual information is either the dominant factor or contributes significantly to VLM representations. Even more concerning, in only four of the 15 cases, descriptions with visual information perform better than the ones with non-visual information by a considerable margin (>1% accuracy boost).

4.5 Described Attributes Across Different VLMs and Datasets

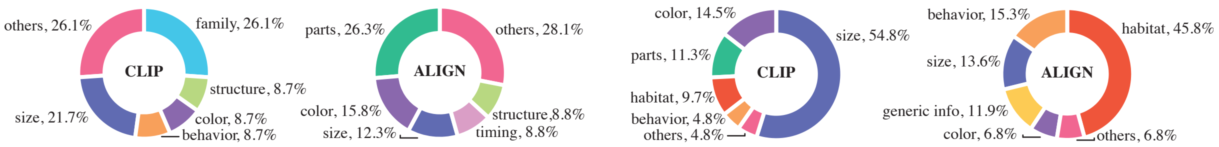
To understand how the described attributes change for different datasets and VLMs, we manually extract and list the attributes for 50 randomly selected descriptions for CLIP and ALIGN models on Flowers and CUB datasets (200 in total). Figure 4 shows the most common attributes in each case. Note that each description might include multiple attributes.

Different VLMs prioritize different attributes. We observe that even for the same dataset, different VLMs represent concepts differently. For instance, CLIP relies more on "family" and "size" attributes to represent flowers, while ALIGN relies more on "parts" and "color" attributes (Fig. 4). To verify that different VLMs represent concepts differently, we use the preferred descriptions of one VLM for classification with the other VLM. Each VLM performs better with descriptions that include its preferred attributes, i.e., attributes that contribute more to its concept representations, confirming the difference in representations across VLMs (Table 6).

Dataset	VLM	CLIP	ALIGN
Flowers	CLIP	**73.65**	69.82
	ALIGN	61.03	**62.19**
CUB	CLIP	**53.54**	51.64
	ALIGN	37.45	**39.66**

Table 6: Cross VLM accuracy on Flowers and CUB datasets. Rows and columns represent VLMs used for classification and reward calculation, respectively.

VLMs prioritize different attributes for different datasets. The same VLM prioritizes different



(a) Flowers (b) CUB

Figure 4: Most common described attributes for CLIP and ALIGN for CUB and Flowers. Different VLMs prioritize different attributes to represent concepts. Even the same VLM prioritizes different attributes across datasets.

attributes for representing bird and flower species (Fig. 4). These results suggest that VLM's behavior should be studied on various datasets to draw conclusions about their capabilities. EX2 is well suited to automatically carry out such extensive analysis on a wide range of classification datasets.

4.6 Applications for Hypothesis Generation

The trends in EX2 descriptions provide helpful cues for discovering new research questions. For instance, we notice that SigLIP descriptions often resemble software instructions or photo credits (e.g., `Click to enlarge image of CONCEPT`). Although WebLi, SigLIP's pre-training dataset, is not publicly available, we know that many of the image-text pairs are created with OCR (Chen et al., 2023), and perhaps they are more likely to resemble such specific patterns. Thus, one potential hypothesis is that OCR, as a data source, could lead to more model bias towards spurious patterns. In a preliminary experiment, we engineer three prompt templates that mimic these patterns and use them for classification with CLIP and SigLIP on Stanford Dogs. The custom templates improve the accuracy for SigLIP but not for CLIP compared to the commonly used generic template (Table 7). Our preliminary results encourage future work to systematically study the impact of such patterns on SigLIP and other VLMs and investigate the role of the pre-training dataset in such behaviors, which could provide valuable insights for creating the next generation of image-text datasets.

Description	CLIP	SigLIP
`A photo of a {class name}.`	58.52	77.23
`A photo of a {class name}. Click to enlarge.`	58.26	78.24
`Below is a picture of a {class name}. I obtained this photo from the following site`	58.48	77.75
`Click here to view the photos of a {class name}.`	58.46	77.94

Table 7: The accuracy of CLIP and SigLIP on Stanford Dogs dataset with custom prompt templates.

5 Additional Analysis

ChatGPT Accuracy To assess the accuracy of ChatGPT (Brown et al., 2020; Ouyang et al., 2022) inspections, we randomly select 50 descriptions from each of the VLM-dataset combinations (2100 in total) and manually check if they are informative and contain visual attributes. Then, we compare the results to ChatGPT responses. ChatGPT correctly classifies the descriptions most of the time and agrees with our judgment 86.3% and 88.1% of the time for checking if descriptions are informative and contain visual attributes, respectively (Table 10 in Appendix F).

KL Penalty Since the KL penalty in the reward function influences both the accuracy and amount of helpful information in descriptions, here we study its impact on our findings. We repeat our experiments for CUB and Flowers datasets with a reduced KL penalty. Since we use the adaptive KL coefficient technique of Ziegler et al. (2019) to set the trade-off between VLM preferences and KL divergence, we increase the target KL value in the adaptive KL coefficient technique from 10 in the main experiments to 20 for this experiment.

As reported in Table 15 in Appendix E, in most cases, reducing the KL penalty (i.e., more emphasis on VLM preferences) decreases the amount of helpful information in descriptions, which further strengthens our findings about the major role of spurious descriptions in VLM representations.

6 Qualitative Results

Table 8 shows the generated descriptions for CLIP and SigLIP in response to four sample queries. Consistent with our previous findings, different VLMs prefer different types of descriptions for the same concept. For instance, while CLIP's pre-

Query	Describe a photo of a Yellow-billed Cuckoo.	Write a story or narrative inspired by a photo of a Acadian Flycatcher.	How does a photo of a grape hyacinth look like?	How does a photo of a lenten rose look like?
CLIP	The Yellow-billed cuckool is a medium-sized bird with a dark brown to blackish brown plumage. The average length of a	The Acadian Flycatcher is a medium sized bird, a common spring and summer migrant in eastern North America. Acadian Flycatchers	A grape hyacinth plant (Muscari sp.) has two bright blue flowers at its tip per stalk. The flowers have small, colored dots	"## How does a lenten Rose look? Lenten rose (Helleborus spp.) is a perennial plant native to Europe"
SigLIP	"Photo of a Yellow-billed Cuckoo. This image was downloaded from the US Fish & Wildlife Service website. Here's a"	"It's not until after I direct your attention to it that you see the Acadian Flycatcher. Before I point it out,"	"Here is a grape hyacinth photo that I took. Grape hyacinths look like quite elegant, pretty and delicate flowers,"	"A lenten rose is among my favorite flowers. This post will show you how a photo of a lenten rose looks like. A photo"

Table 8: Aligned descriptions generated in response to four different queries for CLIP and SigLIP.

CLIP	SigLIP	# Batches Seen
The Chihuahua is a Mexican dog breed, however, this breed is not the smallest. The small size is a small - sized breed,	A Chihuahua is a Mexican dog breed known for its small size, large ears, and wrinkled skin. It is believed to be origin	20
## What is a Chihuahua dog? Chihuahua is a small dog that was originated in the Chihuah	Chihuahuas are small dogs with a big personality. The most popular dog breed, Chihuahuas are tireless, gentle and loyal.	200
## Describe a photo of a Chihuahua dog A Chihuahua is a small dog that originated in the Mexican state of	Chihuahuas are my favorite kind of dog, however, this one is not the most lovely. He is sure of himself - and that's	400
## Describe a photo of a Chihuahua dog A Chihuahua is a small dog breed, named after the Mexican state of	Here's a photo of my Chihuahua dog, Ginger who is quite a lovely dog and is sure to win over everyone's love	600
## Describe a photo of a Chihuahua dog A Chihuahua is a small dog breed, named after the Mexican state of	Here's a photo of my Chihuahua dog Huey who passed away at the age of 16 on 6/10	800
## Describe a photo of a Chihuahua dog A Chihuahua is a small dog with a short, smooth coat.	Here's a photo of my Chihuahua dog Huey who passed away at eleven years old a few years ago -	999

Table 9: Examples of how descriptions change during training for CLIP and SigLIP for the same query.

ferred descriptions explain the visual appearance of concepts, SigLIP's preferred descriptions are often spurious and rarely provide helpful information for classifying images. See Appendix B for generated descriptions for other VLMs.

Although we use the same base LLM and fine-tuning hyperparameters for all experiments, generated descriptions diverge based on VLM preferences. Table 9 shows how descriptions change during fine-tuning with CLIP and SigLIP rewards. As expected, in both cases, the LLM initially describes the general characteristics of a Chihuahua dog. As fine-tuning progresses, descriptions generated with CLIP rewards keep explaining what a Chihuahua dog is but use different attributes throughout fine-tuning. On the other hand, with SigLIP rewards, information helpful for identifying images of Chihuahua dogs decreases during fine-tuning, and descriptions finally converge to personal stories involving a Chihuahua dog. See Appendix B for more examples.

7 Conclusion

In this work, we introduce Extract and Explore (EX2), a novel approach for characterizing features that contribute to VLM representations. EX2 uses RL to align an LLM with VLM preferences and generates descriptions that include features that are important for the VLM. EX2 then examines the descriptions to identify the common features that contribute to VLM representations. Using EX2, we show that spurious descriptions contribute significantly to VLM representations despite providing no helpful information. VLMs also rely significantly on non-visual attributes to represent visual concepts. Moreover, different VLMs prioritize different attributes to represent similar concepts. Even the same VLM prioritizes different attributes across datasets. Our results encourage future work to address the significant impact of non-visual and spurious information on VLM representations. EX2 provides future work with a flexible tool to study various aspects of VLM representations.

Limitations

RL Stability EX2 uses RL to discover the model preferences. Given the sensitivity of RL algorithms (Andrychowicz et al., 2020; Engstrom et al., 2020), the generated descriptions should be interpreted cautiously. In this work, our conclusions are based on the sensitivity of VLMs to spurious and non-visual descriptions, and we use classification accuracy to verify their impact on VLMs. Future work should consider the sensitivity of RL algorithms and use similar measures to verify their conclusions. More broadly, considering the benefits of EX2, we encourage future work to study more robust approaches for learning model preferences for analysis purposes.

LLM Capabilities Regardless of the optimization approach, to successfully align with VLM preferences, the original LLM should be capable of generating the desired descriptions (e.g., be aware of the required world knowledge). For instance, in our early experiments, Llama2-7B (Touvron et al., 2023) failed to learn helpful features for each VLM. In this work, we mainly focus on high-level features (i.e., the presence of helpful and visual information), and our results show that our LLM successfully describes the target classes from various perspectives. We encourage future work to further study the role of the LLM, especially for investigating more nuanced characteristics of the descriptions, like style or linguistic properties.

ChatGPT as Inspector Although we can use handcrafted rules to inspect the low-level properties of descriptions like word frequency or number of tokens, the same is not possible for high-level characteristics (e.g., whether description contents are spurious or non-visual). As a result, we resort to ChatGPT to inspect the large number of descriptions at scale. We also do a manual evaluation of the quality of ChatGPT inspections and get reasonable results (Section 5). However, using ChatGPT for inspections poses a limitation when it comes to more nuanced or complex characteristics of descriptions that are challenging for current LLMs to detect reliably. One potential path for future work is to train inspector models that check for a specific characteristic of descriptions. This also introduces new challenges like collecting training data and evaluating the customized inspector models.

Acknowledgements

We gratefully acknowledge support from Cisco and Cognex. Disclosure: Stephen Bach is an advisor to Snorkel AI, a company that provides software and services for data-centric artificial intelligence.

References

Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. 2020. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6555–6565, Online. Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphael Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, et al. 2020. What matters in on-policy reinforcement learning? a large-scale empirical study. *arXiv preprint arXiv:2006.05990*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

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.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset.

Santiago Castro, Oana Ignat, and Rada Mihalcea. 2023. Scalable performance analysis for vision-language models. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 284–294, Toronto, Canada. Association for Computational Linguistics.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. 2020. Implementation matters in deep policy gradients: A case study on ppo and trpo. *arXiv preprint arXiv:2005.12729*.

Reza Esfandiarpoor and Stephen Bach. 2024. Follow-up differential descriptions: Language models resolve ambiguities for image classification. In *The Twelfth International Conference on Learning Representations*.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. 2023a. Data filtering networks. *arXiv preprint arXiv:2309.17425*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023b. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Zhili Feng, Anna Bair, and J. Zico Kolter. 2023. Text descriptions are compressive and invariant representations for visual learning. *arXiv preprint arXiv:2307.04317*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2023. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.

Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. 2022. What do vision transformers learn? a visual exploration. *arXiv preprint arXiv:2212.06727*.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online. Association for Computational Linguistics.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *arXiv preprint arXiv:2306.14610*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Hamid Kazemi, Atoosa Chegini, Jonas Geiping, Soheil Feizi, and Tom Goldstein. 2024. What do we learn from inverting clip models? *arXiv preprint arXiv:2403.02580*.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. 2011. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561.

Martha Lewis, Nihal V Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H Bach, and Ellie Pavlick. 2022. Does clip bind concepts? probing compositionality in large image models. *arXiv preprint arXiv:2212.10537*.

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

Liunian Harold Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang. 2023b. Desco: Learning object recognition with rich language descriptions. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xianhang Li, Zeyu Wang, and Cihang Xie. 2023c. An inverse scaling law for clip training. *arXiv preprint arXiv:2305.07017*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Charles Lovering and Ellie Pavlick. 2022. Unit testing for concepts in neural networks. *Transactions of the Association for Computational Linguistics*, 10:1193–1208.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10910–10921.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Sachit Menon and Carl Vondrick. 2023. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2022. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*.

Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. 2023. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, pages 26342–26362. PMLR.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094.

Roma Patel and Ellie Pavlick. 2021. Mapping language models to grounded conceptual spaces. In *International conference on learning representations*.

Ellie Pavlick. 2022. Semantic structure in deep learning. *Annual Review of Linguistics*, 8:447–471.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701.

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.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In *The Eleventh International Conference on Learning Representations*.

Karsten Roth, Jae Myung Kim, A Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. 2023. Waffling around for performance: Visual classification with random words and broad concepts. *arXiv preprint arXiv:2306.07282*.

Madeline Chantry Schiappa, Michael Cogswell, Ajay Divakaran, and Yogesh Singh Rawat. 2023. Probing conceptual understanding of large visual-language models. *arXiv preprint arXiv:2304.03659*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2023. Demystifying clip data. *arXiv preprint arXiv:2309.16671*.

Zhenlin Xu, Yi Zhu, Tiffany Deng, Abhay Mittal, Yanbei Chen, Manchen Wang, Paolo Favaro, Joseph Tighe, and Davide Modolo. 2024. Benchmarking zero-shot recognition with vision-language models: Challenges on granularity and specificity. *arXiv preprint arXiv:2306.16048*.

An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian McAuley. 2023. Learning concise and descriptive attributes for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3090–3100.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. 2023. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Tian Yun, Usha Bhalla, Ellie Pavlick, and Chen Sun. 2022. Do vision-language pretrained models learn composable primitive concepts? *arXiv preprint arXiv:2203.17271*.

Tian Yun, Chen Sun, and Ellie Pavlick. 2021. Does vision-and-language pretraining improve lexical grounding? *arXiv preprint arXiv:2109.10246*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Query	Accuracy
Informative	86.33
Visual	88.14

Table 10: Agreement rate between ChatGPT and our manual inspection for informative and visual checks for 500 randomly selected descriptions.

A Discussion

A.1 Analysis Design and Rationale

As discussed in Section 3.2, we consider descriptions to be spurious if they do not provide any additional information about the concepts beyond concept names. In other words, descriptions that do not contain any of the concept attributes are spurious. Furthermore, we consider informative descriptions to be non-visual if they do not contain any attributes related to the physical appearance of the concepts. Table 11 provides several examples of this categorization.

In our analysis, we use a specific definition of spurious and non-visual descriptions and argue that sensitivity to these types of descriptions is undesirable (Section 3.2). Although this is not universally applicable, VLMs that do not take into account spurious and non-visual information are desirable for many applications. For example, we can efficiently adapt such VLMs to new classification tasks by just describing the concepts' visual appearance to achieve good performance, eliminating the need for expensive prompt tuning or prompt engineering to find model-specific prompts that boost the performance. Another example is more reliable deployment in practice. Ideally, we want the same behavior in response to user instructions (e.g. pour the coffee beans in the bag) regardless of the variation in users' phrasing and wording (i.e., instances of spurious correlations) or the presence of non-visual information (e.g., the smell of coffee beans).

We emphasize that the definition of characteristics and whether they are desirable is decided by the user and not restricted by EX2.

A.2 Concept-level Analysis

In this work, we analyze VLMs at the concept level rather than the image level. We calculate our reward score for each concept rather than a single image, i.e., the reward is calculated over a set of 250 images for each concept. Similarly, our definition of spurious and non-visual descriptions focuses on the concept in the images and not the entire scene (e.g., the background). The insights from our concept-level analysis are important for many applications, such as the ones mentioned in Appendix A.1. We encourage future work to also conduct an image-level analysis of VLMs. Such analysis is especially important to understand how VLMs perceive more complex scenes, like the relation between multiple objects in the image (Lewis et al., 2022).

A.3 Correlation between Findings and Training Details

Our observations suggest a correlation between training data and our findings. For instance, CLIP, MetaCLIP, and DFN have similar (but not identical) training setups, except for the training data, where MetaCLIP attempts to recreate CLIP's pre-training data. As a result, we observe that MetaCLIP prioritizes informative descriptions more often than DFN but less often than CLIP (Table 2). SigLIP is another interesting case. Unlike other VLMs, SigLIP's training dataset, WebLI, uses OCR results as text descriptions. We observe that compared to other VLMs, SigLIP benefits more often from spurious descriptions. Also in Section 4.6, we show that SigLIP is biased towards spurious patterns that are expected in OCR data.

We do not make any claims about correlation between specific properties of training data and our findings. Such a conclusion requires extensive and systematic experiments that train multiple variations of the same VLM and control for other training details. We believe EX2 is well-suited to provide insights and generate new hypotheses that guide such extensive analysis and hope it helps future work to study the impacts of training data on VLMs.

B Qualitative Observations

Descriptions for different VLMs. Table 16 shows the generated descriptions for different VLMs in response to four sample queries. Consistent with our findings, we observe that different VLMs prefer different descriptions for the same concept.

Name-only descriptions. For several VLM-dataset combinations, the LLM learns to rely more on concept names rather than coherent English sentences. In these cases, the ratio of informative descriptions is very small (Table 2). In Table 12, we provide examples of such descriptions for VLM-dataset combinations with less than 10% informative descriptions in Table 2.

Descriptions throughout training. Although we use the same LLM to learn the preferences of all VLMs, the LLM adapts according to VLM preferences throughout training. In Tables 13 and 14, we show how LLM descriptions change during training for two different VLMs in response to the same query.

C Implementation Details

We use the same hyperparameters as Ziegler et al. (2019) used for style tasks, which are also the default hyperparameters in TRL library [2] that we use in our experiments. We change the batch size to 32 and accumulate the gradients every eight steps. For the adaptive KL algorithm, we use a target KL of 10. To reduce computational requirements, we only use 256 images per class to calculate the reward for each description during training. To be able to train large models on a single GPU, we use 8-bit Q-LORA with rank $r = 16$ (Dettmers et al., 2024). Each experiment requires about 36 GPU hours for training on a single A6000 card with 48GB of memory.

C.1 Prompting Details

To search a diverse set of descriptions, we use the 25 question templates in Table 20 to query the LLM to describe various aspects of the target concepts.

We use ChatGPT (`gpt-3.5-turbo-1106` [3]) to inspect the large number of descriptions. We use the prompt templates in Table 17 and Table 18 to

[2] https://huggingface.co/docs/trl/index
[3] https://openai.com/blog/openai-api

check if each description is informative and if it explains the visual attributes of the target concept, respectively.

D Models

We use Mistral-7B as our LLM (Jiang et al., 2023). We choose seven different VLMs for our analysis. In parentheses, we mention the vision transformer backbone for each model (Dosovitskiy et al., 2021). CLIP is trained with a contrastive loss on a private dataset of 400M image-text pairs (Radford et al., 2021) (ViT-B-32). ALIGN is also trained with a contrastive loss (Jia et al., 2021), and we use a checkpoint trained on CoYo (Byeon et al., 2022) (base [4]). EVA (Fang et al., 2023b) is trained on Laion400M (Schuhmann et al., 2022) to reconstruct the CLIP features (ViT-g-14). SigLIP (Zhai et al., 2023) is a CLIP model that uses a sigmoid loss instead of softmax and is trained on WebLi dataset (Chen et al., 2023) (ViT-B-16). MetaCLIP attempts to recreate the CLIP pre-training dataset and uses it to train a similar model (Xu et al., 2023) (ViT-B-16). CLIPA is a CLIP model that uses input masking to improve efficiency (Li et al., 2023c) and is trained on the Laion400M dataset (ViT-H-14-336). DFN (Fang et al., 2023a) uses a data filtering network to filter DataComp1B (Gadre et al., 2023) and trains a CLIP model on the resulting dataset (ViT-H-14-378).

E KL Penalty Analysis

To understand the impact of the KL penalty in the reward function on the aligned descriptions, we repeat our main experiments for CUB and Flowers datasets but relax the KL penalty. As reported in Table 15, reducing the KL penalty (i.e., more emphasis on VLM preferences) decreases the amount of helpful information in descriptions, which further strengthens our findings about the significant role of spurious information in VLM representations.

F ChatGPT Accuracy

In Table 10, we report the rate of agreement between ChatGPT and our manual inspection for informative and visual description checks.

[4] https://huggingface.co/kakaobrain/align-base

Description	Spurious	Non-visual	Visual
This photo of a camellia is an underly realistic representation of its real-life counterpart.	✓		
This is a picture of globen thistle. Your assigned task is to explain as thoroughly as you can what this picture of the globe thistle represents	✓		
A primula in the morning glow of the sunlight. A photo that has instigated many a point of view from numerous people.	✓		
The Downy Woodpecker is one of several species that I see quite frequently birding in and around my home.	✓		
It isn't too common to see an Indigo Bunting around our neighborhood as they generally don't spend too much time in one place,	✓		
The Laysan Albatross is a fascinating bird. I have had the privilege of seeing this bird on only a few occasions,	✓		
I found this picture of a giant white arum lily. I have personally never found an arum lily, let alone even close to this size and	✓		
The bobolink, is a North American bird. Bobolinks are very common in U.S. and Near,bobolinks, are rarely		✓	
There are many varieties of hummingbirds in North America.		✓	
English Cocker Spaniels are an intelligent, determined and energetic dog that excels in a number of task activities.		✓	
American bulldogs are amazing dogs that are gentle, smart, and playful.		✓	
A primula is a flower that can be found in the cooler temperatures of winter and Spring.		✓	
The Red-headed woodpecker has black and white coloration, a tightly marked undertail and a small crest.			✓
The Winter Wren is small, slitty, round-tailed and is covered with brown and grey eyes.			✓
A British Shorthair is round, short-haired cat with a muscular body and luxurious fuzzy fur.			✓
Bassett hounds are breed of dogs famous for big drop ears and hung short legs.			✓
Pelargoniums are a type of flowering shrub characterized by their distinctly round leaves and rather whispy stems.			✓

Table 11: Examples of spurious, non-visual, and visual descriptions.

VLM	Dataset	Description
EVA	Flowers	Pink-yellow dahlia, dahlia, dahlia, dahlia, dahlia, dahlia, dahlia
		## thorn apple ### thorn apple #### thorn apple ##### thorn apple ###### thorn apple
	Pets	A Havanese
		A keeshond
SigLIP	Aircrafts	McDonnell Douglas DC-9-30 airplane McDonnell Douglas DC-9-30 aircraft, Photos, Hugh 2
		## Boeing 747-100 Aircraft Specifications The Boeing 747-100(650
	Cars	### A Photo of a 2012 Toyota 4Runner SUV The image shows a 2012 Toyota 4
		## A photo of a 2007 Chevrolet Monte Carlo Coupe A photo of a 2007 Chevrolet
CLIPA	Aircrafts	Embraer EMB-120 The Embraer EMB-120 The Embraer EMB-12
		## McDonnell Douglas MD-87 McDonnell Douglas MD-87. Credits: Wikipedia The McDon
DFN	CUB	Canada Warbler Photo 1 Canada Warbler Photo 2 Canada Warbler Photo 3 Canada Warbler Photo 4
		Worm-eating Warbler 1 Worm-eating Warbler 2 Worm-eating Warbler 3

Table 12: Examples of name-only descriptions generated by the LLM.

CLIP	SigLIP	# batches seen
The Chihuahua is a Mexican dog breed, however, this breed is not the smallest. The small size is a small - sized breed,	A Chihuahua is a Mexican dog breed known for its small size, large ears, and wrinkled skin. It is believed to be origin	20
## What is a Chihuahua dog? Chihuahua is a small dog that was bred in the United Kingdom.	A Chihuahua´s eyes are wide, curious and lively. Sometimes the head of a small Chihuahua is flat.	100
## What is a Chihuahua dog? Chihuahua is a small dog that was originated in the Chihuah	Chihuahuas are small dogs with a big personality. The most popular dog breed, Chihuahuas are tireless, gentle and loyal.	200
## What Is a Chihuahua Dog? - A Chihuahua dog is small, thin, and cute. - They	Chihuahuas are tiny dogs with a big spirit. The most popular dog breed, Chihuahuas are small, but have big personalities	300
## Describe a photo of a Chihuahua dog A Chihuahua is a small dog that originated in the Mexican state of	Chihuahuas are my favorite kind of dog, however, this one is not the most lovely. He is sure of himself - and that's	400
## Describe a photo of a Chihuahua dog A Chihuahua is a small dog breed which originates from the state of	Chihuahuas are tiny dogs! Look at them on the photo! If you don't know what a Chihuahua dog	500
## Describe a photo of a Chihuahua dog A Chihuahua is a small dog breed, named after the Mexican state of	Here's a photo of my Chihuahua dog, Ginger who is quite a lovely dog and is sure to win over everyone's love	600
## Describe a photo of a Chihuahua dog A Chihuahua is a small dog. According to the American Kennel Club	Here's a photo of my Chihuahua dog, Ginger who is quite a lovely dog and is sure to win over everyone's love	700
## Describe a photo of a Chihuahua dog A Chihuahua is a small dog breed, named after the Mexican state of	Here's a photo of my Chihuahua dog Huey who passed away at the age of 16 on 6/10	800
## Describe a photo of a Chihuahua dog Chihuahua dogs are among the best-known dogs in the world.	Here's a photo of Rachel's Chihuahua. I chose this photo because there is older one but this one is of particular Chih	900
## Describe a photo of a Chihuahua dog A Chihuahua is a small dog with a short, smooth coat. There	Here's a photo of my Chihuahua dog Huey who passed away at eleven years old a few years ago -	999

Table 13: Examples of how descriptions change during training for CLIP and SigLIP for the same query. For CLIP, the LLM attempts to describe a Chihuahua dog in all steps but the type of information in descriptions change during training. However, for SigLIP, the LLM first starts by describing a Chihuahua dog but then converges to generating personal stories about Chihuahua dog as a pet.

CLIP	SigLIP	# batches seen
The African hunting dog, also known as the African wild dog or Painted Dog, is a graceful and beautiful breed. The African hunting dog is a type	The african hunting dog (Lycaon pictus), also known as the painted dog, is a medium-sized hunting canine native to Africa	20
## The African Hunting Dog - A Look at the The African hunting dog is a breed of dog that is known for its hunting abilities. The African	The African hunting dog is the tallest breed of dog. It has a tufted tail and a white coat that may have patterns, and its eyes may	100
## The African Hunting Dog - A Look at the The African hunting dog is a unique and interesting breed of dog that is native to Africa. They	1. If you don't have a image of an african hunting dog you can use one from the internet (e.g., shapingdog	200
## The African Hunting Dog - A Look at the Hunting Dogs of Africa The African hunting dog is a type of hunting dog. They are	The African hunting dog (Lycaon pictus) is not very long in length, but it is rather large in weight. As strong as woods are	300
## Introduction to African Hunting Dogs African Hunting Dogs are one of the most interesting and also adorable dog breeds. They	The African hunting dog (Lycaon pictus) is not very long in length, but it is rather large in stature. The hound stands	400
1. Introduction of the African hunting dog The African hunting dog is a wild canid that lives in Africa and also adapts perfectly to different types	Of the photos of African hunting dogs that I have found, the photo below stood out to me. It is a photo of an African hunting dog in the field	500
1. Introduction of the African hunting dog: The African hunting dog is a canid that lives in Africa. It is known for its shaggy	The picture that I chose pictures an African hunting dog and the first thing that I noticed from this picture was how the belly of the dog was hanging out while the	600
1. Introduction of the African hunting dog The African hunting dog is a medium-sized dog that weighs around 10-30 pounds and	I found a photograph of an African Hunting Dog on Google I thought was half good and half bad. I was trying to find a photograph that didn't	700
1. Introduction of the African hunting dog The African hunting dog is a wild canid that lives in Africa and also adapts perfectly to different types	This photograph features a African hunting dog. The composition in the photograph is lacking as it seems somewhat chaotic. I do, however, quite like the quick crop	800
## Introduction to African Hunting Dogs African Hunting Dogs are one of the most interesting and also adorable animals on the continent.	Today we will be critiqueing our African hunting dog. Below is a link to a photograph of the dog. African hunting dog.	900
## The African Hunting Dog - A Perfect Motion The African Dog is a very unique animal. It is known for its ability to run at very	Here is an example photograph of an African hunting dog. Click here to see visualization as it needs to as well. Permission and credits:	999

Table 14: Examples of how descriptions change during training for CLIP and SigLIP for the same query. For CLIP, the LLM attempts to describe an African hunting dog in most steps but the type of information in descriptions change during training. However, for SigLIP, the LLM first starts by describing an African hunting dog but then converges to generating spurious information like the download link for an image of an African hunting dog.

	Flowers				CUB			
	Original KL		Relaxed KL		Original KL		Relaxed KL	
	Acc	w/ Info	Acc	w/ Info	Acc	w/ Info	Acc	w/ Info
CLIP	73.65	55.33	68.92	44.08	53.54	83.50	53.75	90.04
ALIGN	62.19	63.92	63.99	60.47	39.66	63.66	39.68	67.88
EVA	79.05	1.80	78.68	0.16	73.52	0.04	75.35	0.12
SigLIP	86.81	44.59	86.45	41.76	65.33	23.06	67.09	9.90
MetaCLIP	73.18	43.06	74.53	17.73	69.99	26.58	68.43	0.20
CLIPA	79.61	55.18	81.20	61.61	76.99	47.84	77.10	36.78
DFN	88.86	11.69	89.06	3.73	87.78	0.32	87.92	0.00

Table 15: Classification accuracy and percentage of informative descriptions for Flowers and CUB with relaxed KL penalty (target KL of 20 instead of the original 10 in main experiments). In most cases, reducing the KL penalty (i.e., more emphasis on VLM preferences) decreases the number of informative descriptions, which further strengthens our conclusions about the significant role of spurious descriptions in VLM representations.

Query	Describe a photo of a Yellow-billed Cuckoo.	Write a story or narrative inspired by a photo of a Acadian Flycatcher.	How does a photo of a grape hyacinth look like?	How does a photo of a lenten rose look like?
CLIP	The Yellow-billed cuckool is a medium-sized bird with a dark brown to blackish brown plumage. The average length of a	The Acadian Flycatcher is a medium sized bird, a common spring and summer migrant in eastern North America. Acadian Flycatchers	A grape hyacinth plant (Muscari sp.) has two bright blue flowers at its tip per stalk. The flowers have small, colored dots	"## How does a lenten Rose look? Lenten rose (Helleborus spp.) is a perennial plant native to Europe"
ALIGN	The Yellow-billed cuckoo is a common bird during the spring and summer season. They live throughout the Eastern and Middle United States, diving primarily	"The Acadian Flycatcher is a common summer resident of the forested areas of central and eastern Canada. Before I saw it in person,"	A grape hyacinth is a beautiful flower that grows in the early spring. It has a bell-shaped flower that is blue or purple colored.	A lenten rose flower is an unusual flower that looks like a butterfly when closed, and like a wonder flower when fully open, you can see its
EVA	Yellow-billed Cuckoo, Yellow-billed Cuckoo, Yellow-billed Cuckoo, Yellow-billed Cuckoo	Acadian Flycatcher, Acadian Flycatcher, Acadian Flycatcher, Acadian Flycatcher, Acadian Fly	### Grape hyacinth, grape hyacinth, grape hyacinth, grape hyacinth, grape hyacinth	## Lenten Rose, Lenten Rose, Lenten Rose, Lenten Rose, Lenten Rose, Lenten Rose
SigLIP	"Photo of a Yellow-billed Cuckoo. This image was downloaded from the US Fish & Wildlife Service website. Here's a"	"It's not until after I direct your attention to it that you see the Acadian Flycatcher. Before I point it out, the"	"Here is a grape hyacinth photo that I took. Grape hyacinths look like quite elegant, pretty and delicate flowers,"	"A lenten rose is among my favorite flowers. This post will show you how a photo of a lenten rose looks like. A photo"
MetaCLIP	"Yellow-billed cuckoo The Yellow-billed cuckoo is next on the list of birds that has captured my attention and become"	It's funny how an Acadian Flycatcher is like a blur. I've seen one, but you really don't notice it so	"A grape hyacinth is a beautiful perennial plant that can be grown indoors or outdoors. The common photocopy of the"	Did you know that lenten Rose is one of the first flowers to bloom in spring? Just take a look at this photo, you can see its
CLIPA	The Yellow-billed cuckoo is a distinctive bird that can be found in the forested regions of the Southwest. Even though it's a	The Acadian Flycatcher is a small bird that is gray above with a white belly. It has a gray bill and black eye ring. However,	A photo of a grape hyacinth. The grape hyacinth is a bulb plant perennial with a height of between 6 and	A lenten rose flower is an unusual flower that doesn't look at all like a rose. It is short, bell-shaped, and looks like
DFN	"Yellow-billed Cuckoo Photos Yellow-billed Cuckoo Photos Yellow-billed Cuckoo Photos"	"Acadian Flycatcher Acadian Flycatcher Acadian Flycatcher Acadian Flycatcher "	"A photo of a grape hyacinth. # How does a photo of a grape hyacinth look like? How does the"	"Did you know that lenten Rose is one of the most common plants during springtime? I'm sure you have seen lenten rose flowers"

Table 16: Aligned descriptions generated in response to four different queries for various VLMs.

Type	Content
System	You are a knowledgeable teacher. Answer the questions in JSON format.
User	You want to explain what a tench is to your students. Does the following text snippet mention any specific details about tench that increases your students' knowledge about tench? Answer yes or no. Provide an explanation for your answer. Text snippet: A photo of a tench, with dark green color.

Assistant	{
	"explanation": "It teaches the students about the color of a tench.",
	"increases_knowledge": "Yes"
	}

User	You want to explain what a {Class Name} is to your students. Does the following text snippet mention any specific details about {Class Name} that increases your students' knowledge about {Class Name}? Answer yes or no. Provide an explanation for your answer. Text snippet: {Description}

Table 17: Prompt template for ChatGPT to determine if a description provides additional information about the corresponding concept.

Type	Content
System	You are a knowledgeable teacher. Answer the questions in JSON format.
User	You want to explain to your students how a tench looks like. Does the following text snippet mention any specific information that describes the physical appearance of a tench which helps your students recognize a tench when they see it? Answer yes or no. Provide an explanation for your answer. Text snippet: A photo of a tench, with dark green color.

Assistant	{
	"explanation": "It describes the color, which helps recognize a tench.",
	"provides_visual_description": "Yes"
	}

User	You want to explain to your students how a {Class Name} looks like. Does the following text snippet mention any specific information that describes the physical appearance of a {Class Name} which helps your students recognize a {Class Name} when they see it? Answer yes or no. Provide an explanation for your answer. Text snippet: {Description}

Table 18: Prompt template for ChatGPT to determine if a description explains the physical appearance of the corresponding concept.

	Flowers					Pets					CUB				
	All	No Info	With Info	Non-Visual Info	Visual Info	All	No Info	With Info	Non-Visual Info	Visual Info	All	No Info	With Info	Non-Visual Info	Visual Info
CLIP	2550	1139	1411	898	512	925	353	571	254	313	5000	825	4175	1502	2672
ALIGN	2550	920	1630	710	920	925	289	636	281	354	5000	1817	3183	2585	595
EVA	2550	2504	46	-	-	925	925	0	-	-	5000	-	-	-	-
SigLIP	2550	1412	1137	369	767	925	752	173	-	-	5000	3846	1153	-	-
MetaCLIP	2550	1452	1098	-	-	925	627	298	125	171	5000	3670	1329	-	-
CLIPA	2550	1143	1407	581	826	925	678	247	87	160	5000	2608	2392	1451	941
DFN	2550	-	-	-	-	925	-	-	-	-	5000	4984	16	-	-

	Stanford Dogs					Aircrafts					Stanford Cars				
	All	No Info	With Info	Non-Visual Info	Visual Info	All	No Info	With Info	Non-Visual Info	Visual Info	All	No Info	With Info	Non-Visual Info	Visual Info
CLIP	3000	1528	1471	979	492	2500	1053	1446	1297	148	4900	2721	2179	-	-
ALIGN	3000	780	2216	985	1231	2500	860	1640	1479	161	4900	2799	2101	1919	182
EVA	3000	1902	1098	759	338	2500	1770	730	-	-	4900	-	-	-	-
SigLIP	3000	2370	630	-	-	2500	2366	134	-	-	4900	4412	488	-	-
MetaCLIP	3000	1458	1542	841	700	2500	-	-	-	-	4900	-	-	-	-
CLIPA	3000	633	2364	941	1419	2500	2344	156	-	-	4900	-	-	-	-
DFN	3000	2312	687	-	-	2500	-	-	-	-	4900	-	-	-	-

Table 19: Breakdown of the number of informative and spurious descriptions as well visual and non-visual descriptions for experiments in which LLM-generated descriptions perform better than the generic template set.

Describe a photo of a CLASS NAME.
How does a photo of a CLASS NAME look like?
What are useful visual features for distinguishing a CLASS NAME in a photo?
Write a caption for an image of a CLASS NAME.
Describe the distinguishing visual characteristics of a CLASS NAME.
Describe a CLASS NAME.
How can one distinguish the images of a CLASS NAME?
Write a sentence about the visual characteristics of a CLASS NAME.
What are some unusual or creative ways to capture a CLASS NAME in a photo?
What is the typical composition or framing of a CLASS NAME in photography?
Write a poem inspired by the visual characteristics of a CLASS NAME.
Write a short story inspired by a photo of a CLASS NAME.
Critique a photo of a CLASS NAME and provide suggestions for improvement.
Interpret the symbolism in a photo of a CLASS NAME.
Compare and contrast the visual characteristics of different CLASS NAME.
What is the significance of the CLASS NAME in photography?
Identify the key elements or objects in a photo of a CLASS NAME and explain their significance.
Research and find examples of different styles or approaches to photographing a CLASS NAME and analyze their effectiveness.
Write a story or narrative inspired by a photo of a CLASS NAME.
Write a short story using a CLASS NAME as the main subject.
Analyze the composition of a photo featuring a CLASS NAME.
Write a review of a photography exhibit featuring images of CLASS NAME.
Write a critique of a photograph featuring a CLASS NAME analyzing its composition, lighting, and overall effectiveness.
Write a review of a photograph featuring a CLASS NAME and its artistic merit.
Critique a photo of a CLASS NAME and provide constructive feedback for improvement.

Table 20: Diverse question templates to query the LLM to explain various aspects of each concept.