Beyond Accuracy: Metrics that Uncover What Makes a 'Good' Visual Descriptor

Ethan Lin[*1]   Linxi Zhao[1]   Atharva Sehgal[2]   Jennifer J. Sun[1]

[1]Cornell University   [2]University of Texas at Austin

{eyl45, lz586, jjs533}@cornell.edu   atharvas@utexas.edu

Abstract

Text-based visual descriptors—ranging from simple class names to more descriptive phrases—are widely used in visual concept discovery and image classification with vision-language models (VLMs). Their effectiveness, however, depends on a complex interplay of factors, including semantic clarity, presence in the VLM's pre-training data, and how well the descriptors serve as a meaningful representation space. In this work, we systematically analyze descriptor quality along two key dimensions: (1) representational capacity, and (2) relationship with VLM pre-training data. We evaluate a spectrum of descriptor generation methods, from zero-shot LLM-generated prompts to iteratively refined descriptors. Motivated by ideas from representation alignment and language understanding, we introduce two alignment-based metrics—Global Alignment and CLIP Similarity—that move beyond accuracy. These metrics allow us to shed light on how different descriptor generation strategies interact with foundation model properties, offering insights into ways of studying descriptor effectiveness beyond accuracy evaluations.

1. Introduction

Classification-by-description frameworks extend zero-shot classification by pairing each class with more informative text descriptors instead of a simple label. These descriptors—ranging from class names to full sentences—serve as queries to a vision-language model (VLM), which classifies images based on image similarity to each descriptor in a shared embedding space. Such methods have demonstrated improved classification performance and offer greater interpretability by making the model's reasoning process more transparent.

While these approaches have demonstrated success in image classification, a fundamental question remains: What
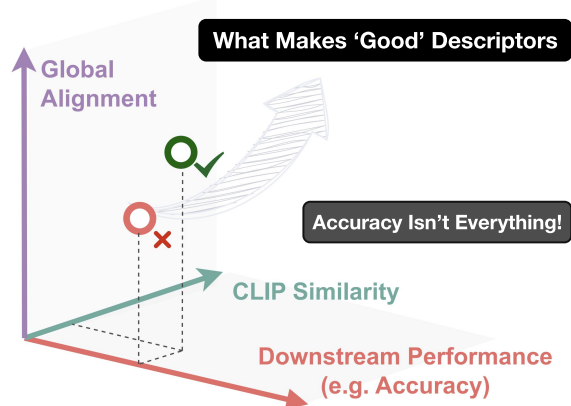

Figure 1. **Overview:** Achieving high accuracy does not guarantee that a set of descriptors is "good". Other factors like interpretability may suffer. Global alignment and CLIP similarity can serve as new metrics for evaluating and understanding different sets of visual descriptors.

constitutes a "good" textual visual descriptor? Descriptors can be obtained in many ways—from using class names directly [19], to leveraging large language models (LLMs) for zero-shot descriptions [14], to employing sophisticated iterative refinement algorithms [21]. Descriptor quality is typically evaluated based on downstream classification accuracy. While informative, this metric offers limited insight into why certain descriptors succeed or fail, and provides little guidance for descriptor discovery [4] or refinement beyond evolutionary search [5] or task-specific feedback loops [21]. Moreover, higher accuracy does not necessarily imply more meaningful descriptors—prior work has shown that appending random words or characters to class names can sometimes improve performance [20]. Clearly, there is need for more principled methods to assess descriptor quality beyond accuracy.

In this work, we propose a novel approach to assessing descriptor quality by probing the relationship between textual descriptors and the underlying VLM. We focus on two key aspects: (1) the representational capacity of a set of descriptors—that is, how well they capture and recon-

---

struct image semantics when used as a projection space; and (2) the relationship between a set of descriptors and CLIP's pre-training data distribution. To study these factors systematically, we analyze a diverse range of text descriptor sets—spanning class names, zero-shot LLM-generated descriptions, and iteratively refined prompts—using two simple but informative metrics (Figure 1). Motivated by recent work on representational alignment [6], we introduce a global alignment metric, $\text{DINO}_{\text{align}}$, which measures the alignment between a vision model's image embedding space and the representation space induced by a set of descriptors using CLIP [18]. We also propose CLIP_{sim}, inspired by findings on language model familiarity with training data [9], which captures how descriptors relate to the CLIP training distribution. Together, these metrics move beyond accuracy to offer a deeper evaluation of text-based visual descriptor quality.

Contributions. This work makes these contributions:
- We propose two novel metrics — *global representation alignment* and *CLIP similarity* — to probe text descriptor quality beyond downstream accuracy.
- We analyze how these metrics correlate with performance across fine-grained classification datasets, and present preliminary analyses that showcase how these metrics characterize the effectiveness of strong text-based visual descriptors.

2. Related Work

Concept Learning with VLMs. Despite rapid improvements in large-scale visual pretraining [2, 7, 11–13, 16, 18, 29, 30], vision-language models (VLMs) still struggle with even moderately complex perceptual questions [3, 8, 24]. Concept learning has emerged as a promising direction, leveraging zero-shot queries to large language models (LLMs) to decompose complex perceptual reasoning tasks into logical combinations of simpler perceptual subtasks – tasks that off-the-shelf VLMs can reliably solve [15, 20, 27, 28]. These logical combinations are typically expressed as learnable linear combinations over an overcomplete set of concepts, sampled from an LLM. Prior work improves this initial concept set either by subsampling or by sampling new concepts using task-specific visual feedback from an LLM [21]. The quality of a set of concepts is quantitatively evaluated using downstream image classification performance. We complement prior work by analyzing the utility of a concept set through three orthogonal lenses: downstream concept accuracy, representational alignment, and alignment with VLM pre-training data.

Representational Alignment with VLMs. Representational alignment refers to the degree of similarity between the internal representations learned by different models

[22]. In the context of VLMs, research has studied both explicit training methods to maximize alignment [7, 18] and the emergence of implicit alignment during training [6, 23].

In this work, we provide the first study of such alignment metrics applied to representations induced by different sets of descriptors. We show how these alignment measures can serve as indicators of the downstream utility and meaningfulness of a concept set.

3. Approach

3.1. Global Alignment Metric

Classification by description using VLMs can be viewed as a semantic projection [26], where images are mapped to similarity scores over a set of textual descriptors. This forms a new representation space, with each dimension corresponding to a specific concept. We assess the quality of the descriptor-induced space using $\text{DINO}_{\text{align}}$, a zero-shot metric that measures the correlation between image similarities in the descriptor space and a reference embedding space (Figure 2). High alignment indicates that the descriptors capture meaningful visual structure.

Following work on the *Platonic Representation Hypothesis* [6], we use a **mutual k-nearest neighbor (M-KNN)** alignment metric to quantify structural similarity between two representations of the same data. Given representation matrices A and B, we compute cosine-based k-nearest neighbor sets K_A and K_B and define the alignment score:

$$\text{Alignment} = \frac{|K_A \cap K_B|}{k}$$

This reflects the average neighborhood overlap between the two spaces. To evaluate the quality and interpretability of a set of text-based visual descriptors, we perform the following steps.
1. Remove class names from all descriptors to avoid bias.
2. Combine descriptors across all classes to form a shared global descriptor set.
3. Extract image embeddings X and concept embeddings Y using a vision-language model (VLM).
4. Project images into the concept space via the image-concepts similarity matrix $S = XY^\top$, where row i represents the image i's similarity with each concept (columns).
5. Compute DINO embeddings Z as the reference space and compute alignment using M-KNN.

3.2. CLIP Similarity Metric

Due to the imbalanced nature of CLIP's training data, semantically similar text descriptors may not yield similar downstream performance. To analyze this, we propose CLIP_{sim}, a metric measuring how well descriptor candidates align with VLM pre-training data (Figure 2). We define
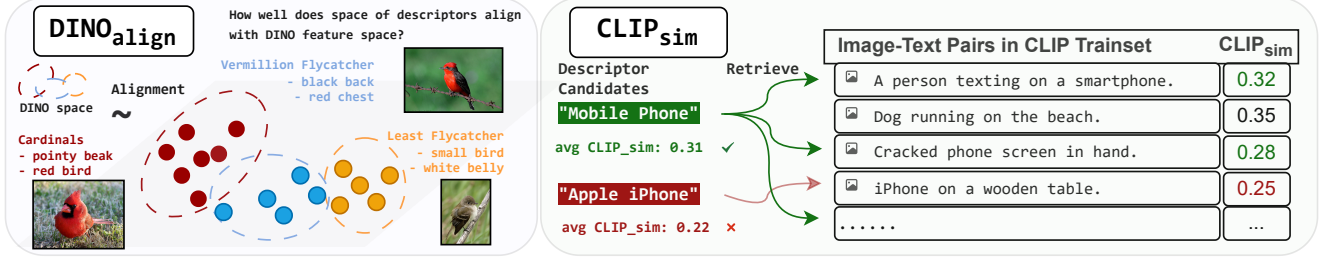
Figure 2. **Overview of our descriptor evaluation metrics.** (**Left**) DINO_align measures how well a set of descriptors preserves class structure by aligning the descriptor-induced similarity matrix with that of DINO image features. (**Right**) CLIP_sim assesses how well descriptors align with visual content by retrieving related image-text pairs from CLIP's pre-training dataset and averaging their similarity scores.

two statistics per descriptor d, using its top-k nearest captions $\{c_1, \ldots, c_k\}$ retrieved by cosine similarity in CLIP's text embedding space. Let $\text{sim}(a, b)$ denote the similarity between two L2-normalized embeddings following Kandpal et al. [9]. Exact or fuzzy string matching is avoided, as it often fails with fine-grained descriptors. We define two statistics per descriptor:

- **Frequency**: number of captions above a similarity threshold τ:

$$\mathcal{I}_d = \{i \in [k] \mid \text{sim}(d, c_i) > \tau\}, \quad \text{Freq}(d) = |\mathcal{I}_d|$$

- **Similarity**: average similarity between each matched caption c_i and its paired image I_i:

$$\text{Sim}(d) = \frac{1}{|\mathcal{I}_d|} \sum_{i \in \mathcal{I}_d} \text{sim}(c_i, I_i)$$

These two metrics captures how frequently a text descriptor appears in CLIP's training set and how well it aligns with visual content in the training set, providing a proxy for how compatible it is with CLIP's vision-language understanding.

Intuitively, high-quality descriptors should not only differentiate between image classes but also align with the inductive biases of vision-language models (VLMs) shaped by their training data. Such descriptors are more likely to yield better downstream performance.

Finally, the overall CLIP_sim score for a method is the average similarity across all descriptors:

$$\text{CLIP}_{\text{sim}} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \text{Sim}(d)$$

4. Experiments

4.1. Setup

Dataset and Model We compute and analyze concept alignment and descriptor frequencies on three image classification benchmarks: CUB-200-2011 [1], NABirds [25], and CIFAR-100 [10]. All experiments are performed using

the test sets of each dataset. These datasets are widely used by previous works and cover both fine-grained and general-purpose categories. We analyze CLIP's pre-training data using the open-source DFN2B dataset.[1] We use CLIP ViT-L/14 [19] to compute zero-shot accuracies and the image-concept similarity matrix. As the current SOTA image embedding space, we compute the reference image embedding using DinoV2-Base [17].[2]

4.2. Descriptor Sets

We evaluate several descriptor sets used for classification:

- *Class Name Prompt*: Uses the standard zero-shot format *"An image of a {class name}"*, popularized by Radford et al. [19].
- *CBD Concepts*: Descriptors from the original Classification by Description framework [14].
- *ESCHER Iterative Descriptors*: Class-specific phrases refined through multiple ESCHER iterations [21].
- *DCLIP (Randomized)*: Inspired by WaffleCLIP [20]; randomly samples a fixed pool of N global descriptors. Each class is formatted as *"{class name}, {descriptor_i}"*.
- *WaffleCLIP Randomization*: Forms descriptors by appending randomized tokens to general concepts and class names, e.g., *"An image of a {concept}: {class name}, which has !32d, #tjli, ˆfs0."*

4.3. Research Questions

Can alignment metrics serve as reliable unsupervised indicators of descriptor set quality or refinement progress? Figure 3 compares accuracy and alignment across datasets. Since alignment is intended to reflect the representational quality of a descriptor set, we remove class names from each descriptor to isolate its effectiveness. The resulting scores follow a consistent pattern:

$$\text{DCLIP} < \text{WAFFLECLIP} < \text{CBD} \leq \text{ESCHER} < \text{CLASS NAME}$$

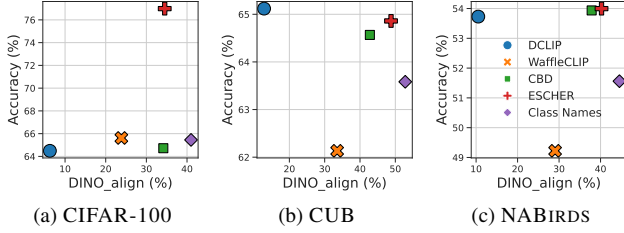(a) CIFAR-100   (b) CUB   (c) NABIRDS

Figure 3. **Classification accuracy versus alignment score** across all datasets. Alignment consistently reflects the quality of descriptor sets, regardless of their classification accuracy.

(a) CIFAR-100   (b) CUB   (c) NABIRDS

Figure 5. We track how descriptor alignment with CLIP's vision-language priors and classification accuracy evolve through iterative refinement.

Viewing classification by description as a form of semantic projection—analogous to PCA—the consistently high alignment scores of class names across datasets suggest that they function like the first principal components: basis directions that capture the highest variance and most distinguishing visual features. When a class-agnostic descriptor set achieves similar alignment, it likely reflects strong semantic grounding with respect to the image space. ESCHER achieves this more reliably than CBD, highlighting the effectiveness of its iterative refinement. In contrast, DCLIP and WAFFLECLIP exhibit much lower alignment despite comparable accuracy, reinforcing that randomly generated descriptors lack meaningful structure. These results suggest that alignment is a consistent and informative signal for assessing semantic quality, even in the absence of class supervision.
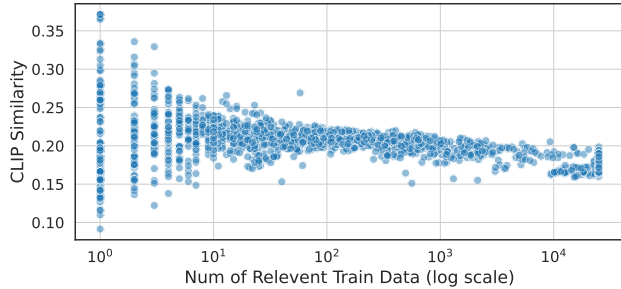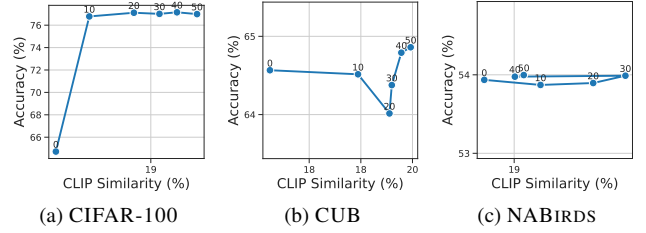

Figure 4. Distribution of descriptors by frequency and corresponding CLIP similarity on NABIRDS. The consistent negative trend suggests that low-frequency descriptors are often more visually grounded and better aligned with CLIP's pretraining space.

Does an iterative algorithm like ESCHER converge toward descriptors that are naturally aligned with CLIP's vision-language priors? As ESCHER iteratively updates text descriptors for a downstream task, we expect the refined descriptors to not only become more discriminative but also better aligned with CLIP's internal representation space. By tracking CLIP_{sim}, we can examine whether ESCHER learns to "speak the language" of CLIP's training data.

To build intuition about CLIP's priors, we first analyze the relationship between descriptor frequency and CLIP similarity. As shown in Figure 4, we observe a consistent negative correlation between frequency and image-text similarity—a somewhat counterintuitive trend. One possible explanation is that frequent descriptors tend to be coarser and more ambiguous (e.g., "animal"), making it harder for CLIP to associate them with a consistent visual pattern. In contrast, rarer descriptors often refer to more specific, visually grounded concepts that CLIP can align with more reliably.

Figure 5 shows how descriptor quality evolves over ES-CHER iterations. On CIFAR-100 and CUB, both CLIP_{sim} and classification accuracy improve steadily, indicating that descriptors become increasingly fine-grained and better aligned with CLIP's representation space. On NABirds, the trend is less pronounced—accuracy remains largely stable across iterations, with a small dip at epoch 40 likely due to noise. These results support our hypothesis: ESCHER's iterative refinement generally improves both CLIP alignment and downstream performance.

5. Conclusion

We present a systematic study of text-based visual descriptors, moving beyond accuracy to examine their alignment with visual features and CLIP's pre-training distribution. Across a range of descriptor types, we find that alignment metrics often correlate with intuitive concept quality—even without access to labels. Notably, iteratively evolved descriptors become better aligned and yield improved downstream performance. These results underscore the value of alignment and frequency analysis as tools for understanding descriptor behavior and foundation model interactions, laying the groundwork for more principled descriptor evaluation and design in future work.

Acknowledgments

References

[1] Technical report. 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning, 2022. arXiv:2204.14198 [cs]. 2

[3] Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation, 2025. 2

[4] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju. Interpreting CLIP with Sparse Linear Concept Embeddings (SpLiCE), 2024. arXiv:2402.10376. 1

[5] Mia Chiquier, Utkarsh Mall, and Carl Vondrick. Evolving interpretable visual classifiers with large language models. 2024. 1, 2

[6] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024. 2

[7] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[8] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? Investigating their struggle with spatial reasoning, 2023. arXiv:2310.19785 [cs]. 2

[9] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge, 2023. 2, 3

[10] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research). 3

[11] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[12] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[14] Sachit Menon and Carl Vondrick. Visual classification via description from large language models, 2022. 1, 3

[15] Sachit Menon and Carl Vondrick. Visual Classification via Description from Large Language Models. 2022. 2, 1

[16] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pages 529–544. Springer, 2022. 2

[17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 3

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, 2021. Number: arXiv:2103.00020 arXiv:2103.00020 [cs]. 2

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and others. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3

[20] Karsten Roth, Jae Myung Kim, A. Sophia Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts, 2023. 1, 2, 3

[21] Atharva Sehgal, Patrick Yuan, Ziniu Hu, Yisong Yue, Jennifer J. Sun, and Swarat Chaudhuri. Self-evolving visual concept library using vision-language critics, 2025. 1, 2, 3

[22] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J. Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment, 2024. 2

[23] Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment, 2025. 2

[24] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578, Seattle, WA, USA, 2024. IEEE. 2

[25] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[26] An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Wang, Jingbo Shang, and Julian J. McAuley. Learning concise and descriptive attributes for visual recognition. *CoRR*, abs/2308.03685, 2023. 2

[27] An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian McAuley. Learning concise and descriptive attributes for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3090–3100, 2023. 2

[28] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification, 2023. 2

[29] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2022. 2

[30] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 2

Beyond Accuracy: Metrics that Uncover What Makes a 'Good' Visual Descriptor

Supplementary Material

6. Experiment Setup

Global Alignment ($\text{DINO}_{\text{align}}$) Details As defined in Section 3.1, $\text{DINO}_{\text{align}}$ quantifies the alignment between the image embedding space of DINOv2 and the representation space induced by a set of descriptors using the **Mutual-KNN** metric. The value of k is chosen based on the average number of images per class of a dataset. This is a simple yet effective heuristic as it generally captures the point where the alignment score stops increasing significantly. Additionally, when computing alignment on a set of descriptors, we remove class names from the descriptors to avoid bias. This is described in more detail in section 8.

CLIP Similarity (CLIP_{sim}) As defined in Section 3.2, this metric quantifies how well a descriptor aligns with CLIP's pretraining distribution. We set the similarity threshold $\tau = 0.7$ and compute statistics over the top 5% of most similar caption-image pairs. To approximate CLIP's training distribution, we randomly sample 5 million image-caption pairs from the DFN2B dataset.[1] Cosine similarity is used for both text-to-text and text-to-image comparisons.

Classification Accuracy We evaluate zero-shot classification performance of a set of descriptors on a dataset using the original classification by description evaluation method [15].

Descriptor Evolution Analysis To study how ESCHER refines descriptors over time, we evaluate two metrics at regular training checkpoints: CLIP alignment (CLIP_{sim}) and zero-shot classification accuracy. All experiments are conducted using the official ESCHER implementation [21].

We perform this analysis on three datasets: CIFAR-100, CUB, and NABirds. For each dataset, we report both metrics at every 10 steps to track the evolution of descriptor quality throughout ESCHER's iterative refinement process.

7. Detailed Results

CLIP similarity across descriptor sets Figure 6 compares CLIP_{sim} across descriptor sets produced by different methods. The scores follow a consistent pattern:

$$\text{DCLIP}, \text{WAFFLECLIP} < \text{CBD} \leq \text{ESCHER}$$

This trend confirms that CLIP_{sim} provides a useful proxy for descriptor quality: descriptors that are more aligned

---

with CLIP's vision-language priors tend to show higher image-text similarity and better downstream performance. While most methods follow this pattern, we observe inconsistent results for CLASS NAME, likely due to their coarse granularity and limited visual specificity.

Descriptor frequency vs. CLIP similarity To further probe how descriptors align with CLIP's training distribution, we analyze the relationship between descriptor frequency (i.e., the number of matching training captions) and CLIP similarity. As shown in Figure 8 and Figure 7, we find a consistent negative correlation across datasets: lower-frequency descriptors tend to achieve higher CLIP similarity. This suggests that rare descriptors may capture more fine-grained, visually grounded concepts that are easier for CLIP to align with, despite being less common in the training set.
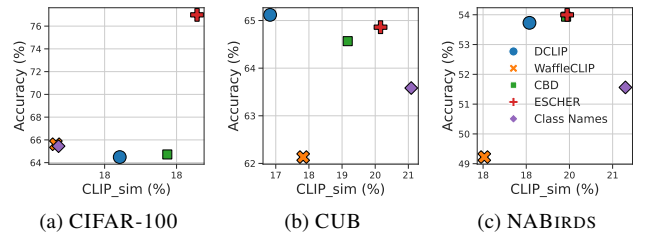

Figure 6. **Classification accuracy versus CLIP_{sim}** across all datasets.
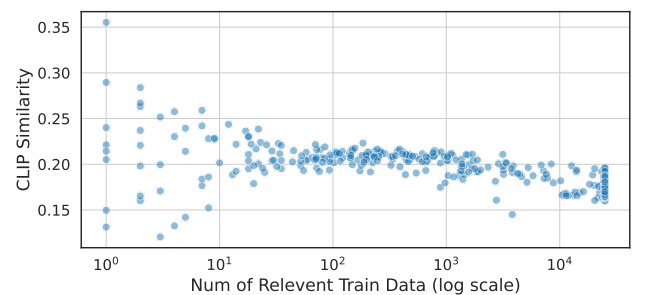

Figure 7. Distribution of descriptors by frequency (number of matching training captions) and corresponding CLIP similarity on CIFAR-100

8. Additional Analysis

Class Name Bias in Alignment In many classification-by-description frameworks—including all those evaluated
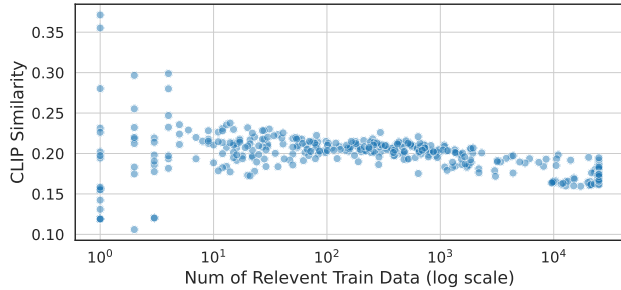
Figure 8. Distribution of descriptors by frequency (number of matching training captions) and corresponding CLIP similarity on CUB

in this work—class names are prepended to each descriptor in the set. For example, descriptors associated with the class *laysan albatross* are formatted as: "laysan albatross, which is a ____". While this improves performance, it also introduces a strong bias that makes it difficult to isolate the quality of the descriptors themselves.

Figure 9 presents alignment and accuracy scores when class names are retained in all descriptors. In contrast to the clearer trends observed when class names are removed, no consistent patterns emerge across methods or datasets—highlighting how strongly class name inclusion can skew alignment metrics. For randomized methods such as DCLIP and WAFFLECLIP, this dependence is expected, as their discriminative ability largely comes from the class name itself. However, the lack of structure in this setting also suggests that even more principled descriptor generation methods may be overly reliant on class name activation, rather than the intrinsic semantic quality of the descriptors.
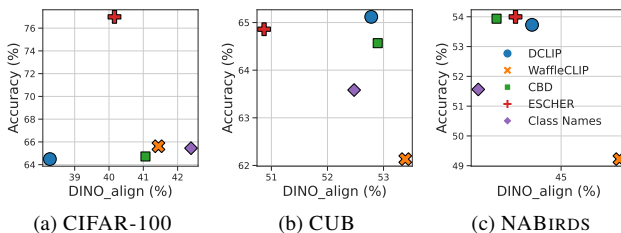

(a) CIFAR-100   (b) CUB   (c) NABIRDS

Figure 9. **Classification accuracy versus alignment score** across all datasets wihtout removing class names from descriptors.

9. Limitations and Future Work

This work presents a preliminary investigation into two novel alignment-based metrics for evaluating the quality of descriptor sets. While we demonstrate their utility across a selected set of descriptor generation methods, our analysis is not exhaustive and opens several avenues for future exploration.

Both metrics have limitations. For $\text{DINO}_{\text{align}}$, computational costs scale with both the number of descriptors and the number of samples, limiting its feasibility for large-scale settings. Moreover, the interpretability of the KNN-based alignment metric likely diminishes when applied to large descriptor sets, and its sensitivity to class imbalance can make cross-dataset comparisons unreliable. As such, alignment is likely most informative when used on compact descriptor sets (e.g., fewer than 100 descriptors), particularly in few-shot settings such as those explored in LM4CV [26] or in evolutionary descriptor search [5], where performance is achieved without reliance on class names.

For CLIP_{sim}, our current analysis is based on a 5M subset of the pre-training data. While this provides a useful approximation, the results could be made more robust by scaling up to a larger portion of the dataset. Future work should explore how descriptor similarity trends generalize across broader samples of the training distribution.

Additionally, our evaluation focuses on a subset of descriptor generation methods. Future studies should incorporate a wider range of approaches, including those that explicitly avoid class name dependence, such as LM4CV [26], or evolutionary search strategies [5]. Applying our metrics to these methods would offer further insight into the quality of different sets of descriptors.

Overall, our metrics lay the groundwork for deeper, label-free evaluations of text-based visual descriptors, and we believe they can help inform the development of more interpretable and effective representations in vision-language models using these descriptors.