# Does VLM Classification Benefit from LLM Description Semantics?

**Pingchuan Ma**[1,2*], **Lennart Rietdorf**[1*],
**Dmytro Kotovenko**[1], **Vincent Tao Hu**[1,2], **Björn Ommer**[1,2]

[1] CompVis @ LMU Munich
[2] Munich Center for Machine Learning

## Abstract

Accurately describing images with text is a foundation of explainable AI. Vision-Language Models (VLMs) like CLIP have recently addressed this by aligning images and texts in a shared embedding space, expressing semantic similarities between vision and language embeddings. VLM classification can be improved with descriptions generated by Large Language Models (LLMs). However, it is difficult to determine the contribution of actual description semantics, as the performance gain may also stem from a semantic-agnostic ensembling effect, where multiple modified text prompts act as a noisy test-time augmentation for the original one. We propose an alternative evaluation scenario to decide if a performance boost of LLM-generated descriptions is caused by such a noise augmentation effect or rather by genuine description semantics. The proposed scenario avoids noisy test-time augmentation and ensures that genuine, distinctive descriptions cause the performance boost. Furthermore, we propose a training-free method for selecting discriminative descriptions that work independently of classname-ensembling effects. Our approach identifies descriptions that effectively differentiate classes within a local CLIP label neighborhood, improving classification accuracy across seven datasets. Additionally, we provide insights into the explainability of description-based image classification with VLMs.

**Code** — https://github.com/CompVis/DisCLIP
**Extended Version** — https://arxiv.org/abs/2412.11917

## 1 Introduction

Human visual recognition is closely related to verbal reasoning, as it often relies on the ability to express visual information in words (Zhao et al. 2022; Shtedritski, Rupprecht, and Vedaldi 2023). However, a neural network usually does not exhibit this property, making its explainability a significant concern for the machine learning community. Some studies (Zhang et al. 2024; Hakimov and Schlangen 2023) aim to connect visual cues and textual descriptions, but these usually require extensive human subject analysis and highly specific datasets with annotations (Young et al. 2014), which are expansive to obtain (Lin et al. 2014).

Vision-Language Models (Radford et al. 2021; Jia et al. 2021) tackle this issue by training neural networks to link images and their textual descriptions within a shared embedding space. This enhances the correlation between visual and textual details. VLMs can be applied to zero-shot image classification by passing an image through the VLM's image encoder and prompting the text encoder with handcrafted inputs like *"a photo of a [classname]"* (Radford et al. 2021). Recent work (Menon and Vondrick 2023) extends this approach by incorporating additional descriptions generated by Large Language Models (LLMs) for each class name. LLMs like GPT-3 (Brown et al. 2020; Ouyang et al. 2022) or Llama (Touvron et al. 2023), trained on extensive text corpora, are intended to provide richer semantics, enhancing VLM classification.

However, *Does VLM classification truly benefit from LLM-generated description semantics?* This work explores this core question, as LLM-generated descriptions present several challenges. For example, descriptions can overlap for similar classes — such as parrots and sparrows — both described as having feathers, which is not a distinguishing feature. Moreover, while supplying the model with as many LLM-generated descriptions as possible may seem advantageous, it results in excessively lengthy collections. This complicates understanding the contribution of each description to the final decision.

Another problem is the structured noise ensembling phenomenon (Roth et al. 2023): LLM-generated descriptions can be replaced with high-level concepts and random characters (such as "*Baklava*", "*a food that is 34mfqr5*") while still improving the model performance. These slightly modified duplications act as test-time augmentation for the original prompt, resulting in an averaged robust output. This raises the question whether the improvement is due to additional semantics of the LLM-generated descriptions or to the ensemble effect of the noise augmentation.

Given these challenges, a proposed model should meet three criteria: 1) As humans who describe with a limited set of descriptions, the model should also operate with a manageable number of text descriptions. 2) These descriptions should be semantically meaningful. 3) The model should be resilient to noise ensembling. To address the issue of noise ensembling, we constrain the model to use only textual descriptions that do not contain the classname, *i.e*. classname-

*These authors contributed equally.

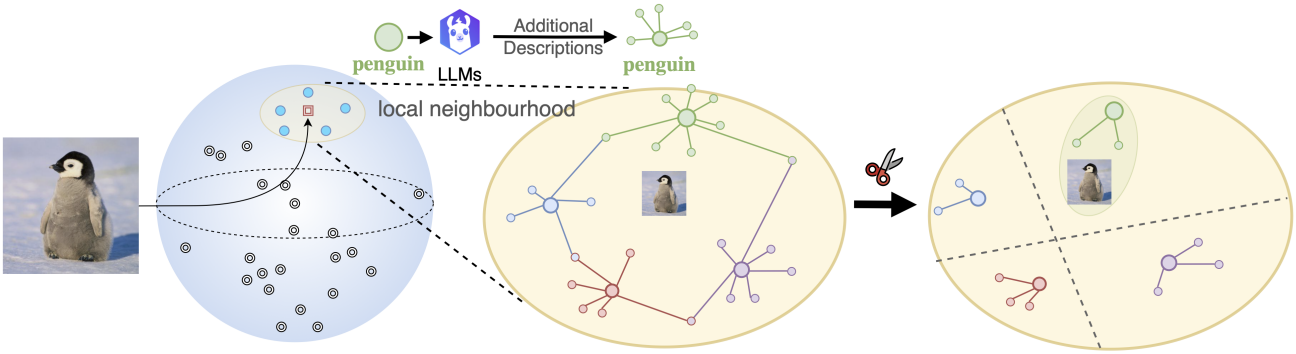



Figure 1: Are the extra semantics provided by LLM truly useful? Our method first identifies candidate labels using only the class name. We then filter out descriptions that may seem logical but do not differentiate the group, *e.g.*, ambiguous, overly generic, or noisy descriptions. This refinement ensures that the remaining descriptions provide distinctive vision-language cues within the local candidate neighborhood, offering more specificity than the class name alone can capture.

free descriptions.

Additionally, we employ a training-free algorithm that processes textual description embeddings within the neighborhood of the queried image embedding. This approach narrows the focus to descriptions relevant to distinguishing between the specific subset of ambiguous classes, reducing the information to process and targeting a more manageable problem rather than attempting to differentiate across all classes. Our method passes the image through the CLIP image encoder, identifies a set of ambiguous class names representing possible candidates, and then applies a straightforward procedure to determine the most distinctive descriptions for these candidates, as shown in Figure 1.

Moreover, our method uses the text embedding of the classname only once and subsequently leverages classname-free LLM-generated descriptions. Therefore, we ensure that performance gains are not due to noisy augmentations of the classnames but rather to a semantically meaningful enrichment. In summary, our contribution is threefold:

- We propose an alternative evaluation scenario for VLM classification tasks to assess whether performance gains stem from genuine semantic understanding rather than an ensemble effect, which is difficult to discern under conventional setups (Table 1 and Figure 3).
- Using this alternative setup, we introduce a training-free approach (Section 3.2) that narrows the focus to a small neighborhood and selects precise, semantically meaningful, and distinguishing class descriptions to improve the VLM classification performance (illustrated in Figure 1).
- Our method achieves improved performance compared to related approaches in two different setups, offering insight into the explainability of fine-grained image classification with VLMs (in Section 4.2).

## 2 Related Works

**Vision-language Models for Classification** Vision-language models such as (Radford et al. 2021) can be used for classification. Notable training-free approaches that build on top of this include DCLIP (Menon and Vondrick 2023) and CuPL (Pratt et al. 2023), where class name texts are augmented with knowledge contained in LLMs to leverage seemingly discriminating characteristics to achieve performance boosts. As Roth et al. (2023) demonstrated, similar effects could be obtained by augmenting the class name with text noise and high-level concepts, raising concerns that many performance gains from DCLIP (Menon and Vondrick 2023) were not due to additional semantics but rather to introduced noise. FuDD (Esfandiarpoor and Bach 2024) introduced contrastive zero-shot prompting to obtain a more diverse set of text prompts. The disadvantage of these approaches is that they rely on ensembling the description extended class name multiple times to achieve significant gains, making it difficult to separate additional semantics from random augmentations of the class name.

Notable approaches that train in the CLIP embedding space include Yan et al. (2023), where nearest neighbors from a pool of text embeddings replace linear weights of a learned dictionary, and LaBo (Yang et al. 2023), which trained a linear classifier on a wide and global bottleneck of language activations selected for diversity and coverage. Zhou et al. (2022) performed non-explainable tuning of text prompt embeddings to optimize classification. In contrast, while Zang et al. (2024) trained the last layer of image and text encoders over a concept bottleneck to discover explainable concepts. Feng, Bair, and Kolter (2023) trained a sparse logistic regression over a matrix of image-language activations, with the training signal also used to train the image encoder.

In contrast to the methods outlined above, our approach delivers humanly understandable, semantically meaningful, disjoint, and distinguishing language descriptions in text space through a training-free method, which boosts VLMs classification accuracy while providing higher explainability. We further discuss the better explainability in Section 3.2.

**Test Time (noise) Augmentation** Data augmentation involves increasing the diversity of training examples without explicitly collecting new data. It can also be employed at test time to enhance robustness (Cohen, Rosenfeld, and Kolter 2019) and improve accuracy (Szegedy et al. 2015; Jin et al. 2018). Notably, simply adding noise to the input string at different levels (Kobayashi 2018; Şahin 2022; Belinkov and Bisk 2018) or their textual embeddings (Sun et al. 2020; Chen, Yang, and Yang 2020; Hao et al. 2023), can achieve similar effects on both performance and robustness across various tasks and domains (Feng et al. 2021).

Test Time Augmentation TTA introduces an *ensemble* of predictions from several transformed or distorted versions of a given test input to obtain a "smoothed" prediction. For example, one could average the predictions from various modified versions of a given string, ensuring that the final prediction is robust to any single unfavorable version (Roth et al. 2023; Menon and Vondrick 2023; Esfandiarpoor and Bach 2024).

Some previous methods (Esfandiarpoor and Bach 2024) used up to hundreds of thousands of descriptions per class, achieving significant improvements in classification accuracy with VLM. However, it is challenging to determine if the performance gains result from the vast ensemble or true information, hence hindering explainability.

# 3 Method

First, we introduce the conventional task formulation for image classification using Vision-Language Models. We then present our unique approach to this task to enable explainability. Finally, we propose a specific solution to enhance the results further.

## 3.1 Background

**VLM for Visual Classification** The process of image classification by Vision-Language Models occurs as follows: Given an image $x$ and a set of class labels $\mathcal{C}$, one classifies the image $x$ by retrieving the class label $\tilde{c}$ with the highest vision-language score:

$$\tilde{c} = \arg\max_{c \in \mathcal{C}} s(c, x), \tag{1}$$

where the vision-language scores $s(c, x)$ use a function $\phi(\cdot, \cdot)$, to calculate similarity scores for image-text embedding pairs. A typical instance of $\phi(\cdot, \cdot)$ is the usual cosine similarity. Traditionally, a vision-language score is obtained in the following way: using the image-text embedding function $e$ of the VLM and given a text representation (a string containing the class name) $t_c$ of class label $c$:

$$s(c, x) = \phi(e(t_c), e(x)), \tag{2}$$

where $e(\cdot)$ is the image or language embedding. Another way to obtain vision-language scores is via ensembling. The motivation for ensembling can be derived from how a human describes an object. For example, when describing an apple, we can describe it as a *"green stuff"*, *"a round object"*, or *"fruit of the same size as an orange"*.

In this case, there is no single text representation $t_c$ for a class $c$ but a set of language representations $\mathcal{D}(c)$ where the ensembling happens over the elements of $\mathcal{D}(c)$:

$$s(c, x) = \frac{1}{|\mathcal{D}(c)|} \sum_{d \in \mathcal{D}(c)} \phi(e(d), e(x)), \tag{3}$$

In the course of this work and w.r.t. to textual descriptions, we call a set $\mathcal{D}(c)$ a description assignment. Furthermore, an LLM "assigns" descriptions by generating them when prompted for a particular class $c$, yielding the assignments $D(c)$. The elements of $\mathcal{D}(c)$ can be pure text augmentations, *e.g.*, *"an image of* `[cls]`*"*, *"a photo of* `[cls]`*"*, or can contain LLM-generated text descriptions and high-level-concepts, *e.g.*, *"an image of* `[cls]`*, a type of* `[LLM-generated category]`*, with* `[LLM-generated descriptions]`*"*. Most of the approaches (Menon and Vondrick 2023; Pratt et al. 2023; Esfandiarpoor and Bach 2024; Roth et al. 2023) that use ensembling as in Equation (3) with LLM-generated contents always include the class name token `[cls]` for $\forall d \in \mathcal{D}(c)$, which we denoted as *classname-included descriptions*.

## 3.2 Our Approach

**Classname-free Descriptions** In the conventional setup (Menon and Vondrick 2023; Pratt et al. 2023; Esfandiarpoor and Bach 2024), performance gains may result from the noise augmentation of the class name text `[cls]` embedding through its various combinations with `[LLM-generated category]`, `[LLM-generated descriptions]`, and even random strings. While random strings should not contribute extra semantics and are likely embedded far away from `[cls]`, this can sometimes apply to LLM-assigned text due to the vocabulary discrepancy between VLMs and LLMs. Another cause may also be that the images do not exhibit the described property. Despite this, such combinations can still perform well, although the assigned descriptions are not semantically correct.

To better investigate whether the improved performance stems from semantic enrichment or the ensemble effect, we propose an approach where, out of all elements in $\mathcal{D}(c)$, exactly one element should contain the class name $c$. The remaining elements must contain textual descriptions without the class name. This set of descriptions then becomes:

$$\mathcal{D}(c) = \{d^{c+}, d_0^{c-}, ..., d_m^{c-}\}, \tag{4}$$

where $d^{c+}$ denotes that the description contains the class name, while $d^{c-}$ denotes that it does not. A typical $\mathcal{D}(c)$ can therefore be the following: {*"An image of apple pie."*, *"crispy brown crust"*, *"graham cracker crust"*}. Whereas in the conventional setup, the `[cls]` would be the following {*"An image of apple pie."*, *"An image of apple pie with crispy brown crust"*, *"An image of apple pie with graham cracker crust"*}. The comparison of different setups is shown in Figure 2.

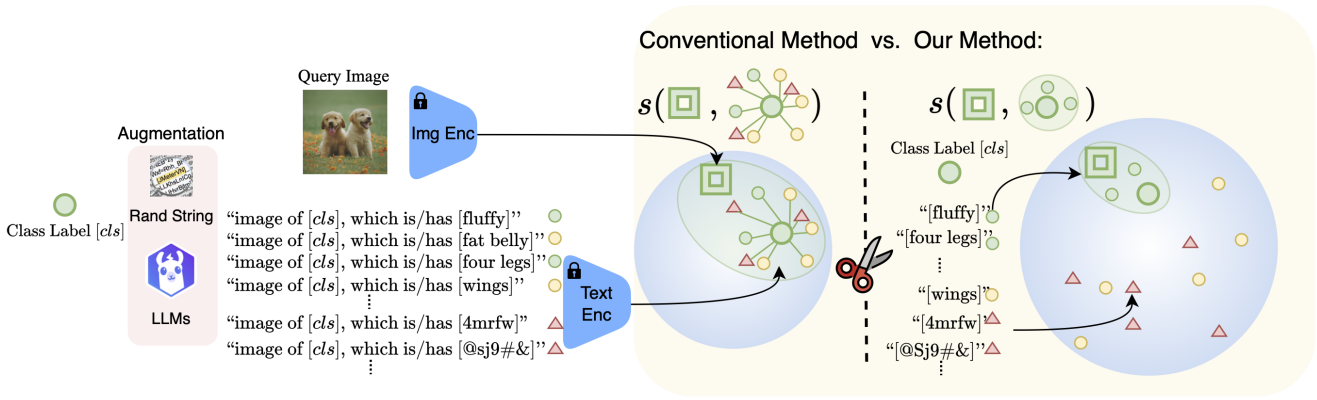



Figure 2: In the conventional setup (*left*), using CLIP with LLM-assigned class descriptions or even random strings can sometimes result in performance gains due to the added semantics or the smoothing ensemble effect. However, when the classname is removed, *i.e.* under the proposed classname-free setup (*right*), these descriptions will fail to perform well, as only meaningful descriptions w.r.t. the class are useful. In contrast, random strings or non-informative descriptions bring no gain.

**Different Weights for `cls`** As discussed previously, the language-ensembling VLM method is evaluated under the conventional scenario by averaging the aggregated class-specific similarities between the images and class-specific descriptions.

However, in our classname-free setup, it is unclear if plain averaging across the obtained classname-free descriptions and the single `cls` is appropriate. This is because the classname is probably the most important text representation of the class, whereas the classname-free descriptions rather have a supporting, distinguishing character. To address this challenge in our evaluation, a weighting factor $w_{cls} \in \mathbb{R}^+$ gets introduced to the vision-language ensemble:

$$s(c, x) = \frac{1}{|\mathcal{D}(c)|} \sum_{d \in \mathcal{D}(c)} w(d) \cdot \phi(d, x) \tag{5}$$

with

$$w(d) = \begin{cases} w_{cls} & \text{if } d = d_{cls}, \\ \frac{1}{|\mathcal{D}(c)|-1} & \text{if } d \in \mathcal{D}(c) \setminus \{d_{cls}\}. \end{cases}$$

Weights of the classname-free descriptions are normalized to one to have the same relative weightings between classes with different amounts of assigned descriptions. Nonetheless, the challenge remains how to find class-specific, classname-free descriptions that actually improve the classification accuracy. This we shall discuss next.

**Selection of Descriptions** Our method works in a local candidate neighborhood: Given a test image $x_i$, one retrieves its top-$k$ predictions based solely on text embeddings of texts such as *"a photo of `[cls]`"*. These preliminary candidate labels constitute the image's local label neighborhood $\mathcal{A}(x_i) = \{a_0, a_1, \ldots, a_k\}$, in which more fine-grained descriptions can offer further distinctiveness.

The image-language similarities of class descriptions can correlate positively or negatively with ambiguous candidate classnames of a test image. Ideally, one wants to find descriptions that only correlate positively with one of the ambiguous classnames and negatively with all the others - hence providing a distinctive and explainable language representation. Consequently, the assignment of classname-free descriptions of classes denoted as $\mathcal{D}(c)$ can significantly influence the final classification result. For example, an albatross might be best distinguished from a penguin by *"sailing through the air"* while it might not be well told apart by *"is a seabird"* since both classes share this feature. Furthermore, this connection must also be well represented in the VLM embedding space.

Algorithm 1 depicts our proposed procedure to find such descriptions. Having available $n$ reference image samples per class $c$ and a global, classname-free description pool $\mathcal{P}$ to select from, the goal is to find a set of descriptions $\mathcal{D}(c) \subset \mathcal{P}$ with $|\mathcal{D}(c)| = m$ that distinguishes each class $a \in \mathcal{A}(x_i)$ from its most ambiguous classes $a' \in \mathcal{A}(x_i) \setminus a$, *i.e.* the small neighborhood of classes around the given images, as depicted in the left part of Figure 1. For that, one utilizes a lookup matrix $S$ containing classwise averaged image-description similarities to obtain feedback from the VLM embedding space. The criterion for assigning descriptions $D(a)$ is a score that is positive if a description activates on average higher for $c = a$ than for all $a' \in \mathcal{A}(x_i) \setminus a$, *c.f.* line 4 of Algorithm 1. This yields $S_i^+$, a positive subset of the lookup similarity matrix $S$.

As $|S^+| > m$ and one wants to extract the most distinctive descriptions from it, the selection heuristic $\Phi$ gets applied to $S^+$. It selects top-$m$ scoring descriptions via:

$$\texttt{top-}m(\mathit{mean}(S^+, dim = 0)), \tag{6}$$

*i.e.*, those $m$ descriptions whose averaged image-description similarity differences to $c$ are, on average, maximally large. In other words, these descriptions activate highly for $a$ but not highly for any $a' \in A \setminus a$ on average. Because these $m$ descriptions are selected without a prepended class name *cls*, they can serve as classname-free language representations of class $c$.

The selected descriptions can then be used as described in Section 3.2 or Section 3.1 for inference. In both cases, clas-

Algorithm 1: Inference: Obtain distinctive language descriptions with feedback from VLM space.

**Require:** $x_i$ - Query image to be evaluated
$\mathcal{P}$ - global description pool obtained from previous stage
$\mathcal{I}$ - probing image embeddings containing few $n$ samples $\forall c \in \mathcal{C}$ in training split
$\mathcal{A}_i$ - a set containing $k$ preliminary labels using standard CLIP retrieval with only $cls$
$\Phi_m$ - selection heuristic to get $m$ descriptions for $\mathcal{A}_i$ from the pool

**Ensure:** output a set of distinctive language descriptions $\mathcal{D}_i \in \mathbb{N}^{k \times m}$

$S \leftarrow \text{matmul}(\mathcal{I}, \mathcal{P}).\text{reshape}(n, |\mathcal{C}|, |\mathcal{P}|).\text{mean}(\text{dim} = 0) \in \mathbb{R}^{|\mathcal{C}| \times |\mathcal{P}|}$ ▷ *Look-up similarity matrix*

$\mathcal{D}_i \leftarrow \{\}$

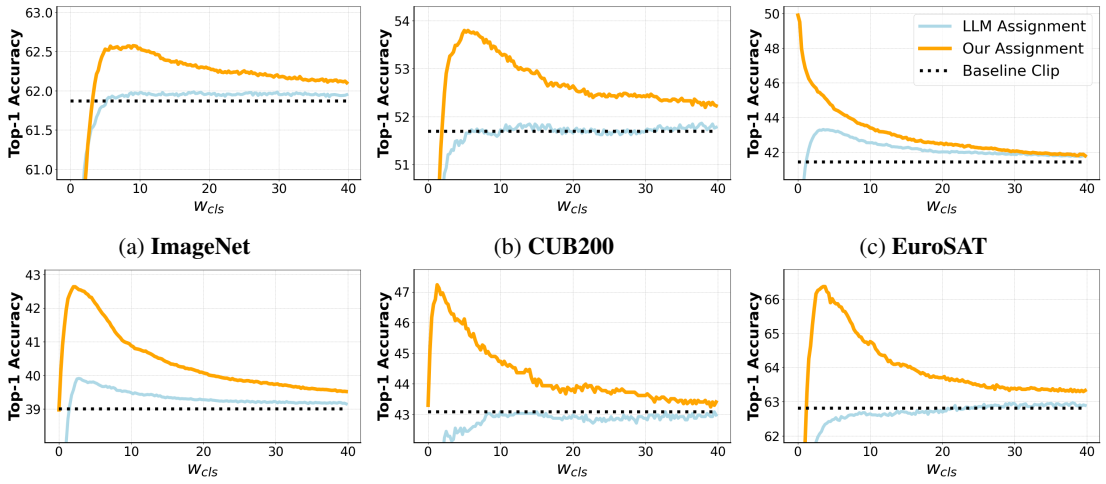
**for** each element $a \in \mathcal{A}_i$ **do**

  $S_i^+ \leftarrow [S[a,:] - S[\mathcal{A}_i \setminus a, :]]^+$ ▷ *Select the positive subset*

  $\mathcal{D}_{i,a} \leftarrow \Phi_m(S_i^+) \in \mathbb{N}^m$ ▷ *Extract $m$ descriptions that distinguish $a$ from the other $\mathcal{A}_i$*

  $\mathcal{D}_i \leftarrow \mathcal{D}_i \cup \mathcal{D}_{i,a}$ ▷ *Descriptions to differentiate $x_i$ from the $k$ preliminary labels.*

$s(\mathcal{D}_i, x_i)$ ▷ *Compute similarity within the local neighborhood*




(a) **ImageNet** (b) **CUB200** (c) **EuroSAT**

(d) **Places365** (e) **DTD** (f) **Flowers102**

Figure 3: Overall Performance of all datasets in classname-free setup. For descriptions assigned by our method and an LLM, $w_{cls}$ assesses the influence of class labels on the performance across different datasets. For a detailed discussion, see Section 4.2.

sification happens via an ensembling (classname-containing and classname-free) of image-language similarities, as introduced in DCLIP. Applying $\arg\max$ over the ensembled, description-enriched image-language similarities of the candidate set $\mathcal{A}(x_i)$ yields the final classification decision of the image.

**Better Explainability** Our proposed method achieves *better explainability* by offering these four characteristics:

1. **The original CLIP encoders for text and image are retained,** rather than fine-tuned to represent a different embedding space, as seen in some works (Zang et al. 2024; Feng, Bair, and Kolter 2023). Hence, our approach preserves the general validity of the CLIP embeddings.
2. **The number of resulting textual descriptions for a single class is kept within a reasonable limit,** similar to the approach in the seminal work (Menon and Vondrick 2023). This helps minimize the potential for noise augmentation, unlike methods that generate hundreds and thousands of descriptions (Esfandiarpoor and Bach 2024).
3. **The overlap between concepts across various classes is minimized,** in comparison to methods with global concept bottlenecks Yang et al. (2023); Yan et al. (2023); Zang et al. (2024). Sparse overlapping ensures clearer distinctions between classes.
4. **We do not use continuous weights over resulting textual descriptions,** as done in Yang et al. (2023); Yan et al. (2023); Zang et al. (2024). Long vectors of continuous weights can be less interpretable compared to clear, discrete indicators of whether a concept is present. Hence, our method offers improved clarity and explainability.

## 4 Experiments

This section evaluates our approach on seven widely used benchmark datasets for (fine-grained) visual classification. We compare our approach to state-of-the-art methods and provide qualitative results.

| Source of $\mathcal{P}$ | Description Assignment | Max #desc. ↓ | ImageNet | ImageNetV2 | CUB200 | EuroSAT | Places365 | DTD | Flowers102 |
|---|---|---|---|---|---|---|---|---|---|
| DCLIP | LLM (global eval) | 13 | 61.99 | 55.09 | 51.79 | 43.31 | 39.91 | 43.09 | 62.97 |
| DCLIP | LLM (local-k eval) | 13 | 61.99 | 55.06 | 51.83 | 43.29 | 39.87 | 43.09 | 62.86 |
| DCLIP | *Ours* | 5 | <u>62.57</u> | **55.48** | **53.80** | **49.89** | <u>42.64</u> | **47.23** | <u>66.37</u> |
| Random | *Ours* | 5 | 62.18 | <u>55.22</u> | 52.31 | 40.82 | 40.44 | 44.73 | 66.12 |
| Contrastive | LLM | 40 | 62.03 | 54.88 | 52.24 | 46.97 | 40.37 | 44.41 | 62.90 |
| Contrastive | *Ours* | 5 | **62.78** | **55.48** | <u>53.45</u> | <u>49.47</u> | **42.65** | <u>46.97</u> | **67.07** |

Table 1: Image classification in classname-free setup with different assignments and pools. Our method consistently produces the highest accuracies in this setting. We use the best-performing $w_{cls}$ of the respective assignment to ensure a fair comparison.

### 4.1 Implementation Details and Datasets

**Implementation Details.** We use CLIP (Radford et al. 2021) as the base Vision-Language Model (VLM) for our approach. Unless stated otherwise, the backbone for CLIP is the ViT-B/32 (Vaswani et al. 2017; Dosovitskiy et al. 2021). We randomly sample a subset from each dataset's standard training split to obtain the lookup similarity table *S* (details see Appendix A.9). Our empirical tests confirmed that this sampling process does not significantly impact performance. The Large-Language Model (LLM) generated descriptions are sourced directly from DCLIP (Menon and Vondrick 2023) or generated using the contrastive prompting method with `gpt-3.5-turbo-1106` and `Llama-3-70b-chat-hf` via APIs.

**Datasets.** We evaluated our methods on the following standard datasets using the standard protocol (classification accuracy) based on previous works (Menon and Vondrick 2023; Roth et al. 2023): ImageNet (Deng et al. 2009), ImageNetV2 (Recht et al. 2019), CUB200-2011 (Wah et al. 2011) (fine-grained bird classification), EuroSAT (Helber et al. 2017) (satellite image recognition), Places365 (Zhou et al. 2017), DTD (Textures, (Cimpoi et al. 2014)), and Flowers102 (Nilsback and Zisserman 2008).

**Source of Obtaining Description Pool $\mathcal{P}$.** These descriptions can be obtained in the following ways: 1) directly from the published descriptions of other works, such as DCLIP (Menon and Vondrick 2023) or FUDD (Esfandiarpoor and Bach 2024); 2) generated based on the provided procedures and code bases of other works, if descriptions are not available; 3) or created through contrastive prompting, which aims to extract meaningful descriptions by contrasting hard negative samples within a neighborhood. The motivation is similar to FuDD (Esfandiarpoor and Bach 2024), but we use significantly fewer descriptions per class. As this is only an alternative for constructing a description pool and orthogonal to our proposed method, we provide more details in Appendix A.7 on the construction of the contrastive pool.

### 4.2 Experimental Results

**Classname-free Evaluation.** We evaluate the quality of the classname-free description assignments selected by our method in the classname-free evaluation setup (cf. Section 3.2). Examples of selected descriptions can be found in Appendix A.12. Performance of our algorithm across 7 classification benchmarks is shown in Figure 3, highlighting how varying $w_{cls}$ impacts top-1 accuracy. The non-ensembled CLIP baseline performance, independent of $w_{cls}$, is also included for reference. Our selected assignments consistently outperform the DCLIP LLM assignments. Notably, for the EuroSAT, Flowers102, CUB200, DTD, and Places datasets, optimal performance occurs when $w_{cls}$ is low ($[0, 10]$), emphasizing the importance of classname-free descriptions while exceeding the baseline performance by up to $9\%$ and the LLM performance by up to $8\%$. However, the LLM-assigned descriptions cannot produce performance gains in the classname-free scenario that comes close to our selections.

Further increasing $w_{cls}$ and thereby weighting the single classname-included description higher reduces accuracy, showing that overly prioritizing the classname diminishes the benefits of our classname-free descriptions.

Interestingly, the smaller gain for ImageNet ($\approx$ 0.5pp.) also corresponds to a lower bump for low $w_{cls}$ in the plot. This may be due to the noisier backgrounds of this dataset, which hinders the selection of generally valid descriptions.

Quantitative results are shown in Table 1, where we report the peak accuracy for each dataset regardless of $w_{cls}$. Interestingly, only 5 selected classname-free descriptions per preliminary class of an image are enough to surpass the performance of the DCLIP LLM assignments. An additional classname-free performance of up to 6.6 pp. (for the EuroSAT dataset) can be achieved.

To confirm that our gains are not driven solely by the image-wise top-$k$ neighborhood, we also evaluate DCLIP LLM assignments in the local top-$k$ context, which shows no significant improvement. This suggests that our approach succeeds by the selection procedure within the top-$k$ neighborhood rather than the search space restriction alone. Importantly, these gains are independent of a specific description pool $\mathcal{P}$ as they also hold for a contrastive prompting pool.

To our knowledge, no prior work has explored a comparable classname-free evaluation setup to determine the true distinctiveness of assigned descriptions $d_0^{c-}, ..., d_m^{c-}$ in combination with a classname prompt $d^{c+}$. However, some works use methods like trainable bottleneck classifiers (Yang et al. 2023; Yan et al. 2023) or embeddings (Zang et al. 2024), which can be considered "classname-free." Despite this similarity, they are too different to compare against (detailed discussion in Appendix A.8).

| Description Assignment | ImageNet | ImageNetV2 | CUB200 | EuroSAT | Places365 | DTD | Flowers102 |
|---|---|---|---|---|---|---|---|
| LLM assignments | 11.65 | 10.69 | 3.47 | 28.11 | 21.36 | 17.77 | 3.19 |
| random assignments | 0.08 | 0.06 | 0.43 | 11.61 | 0.11 | 2.45 | 1.01 |
| *Ours* | **50.16** | **43.98** | **41.53** | **43.24** | **36.36** | **43.09** | **51.52** |

Table 2: Performance in classname-free setup with $w_{cls} = 0$. Our descriptions are robust and perform well, even if the classname text $D_{cls}$ is weighted by $w_{cls} = 0$. LLM assignments give a considerably worse performance in this scenario. Randomly assigned descriptions fail to provide reasonable guidance. Llama3-70B with Contrastive Prompting is used as source pool $\mathcal{P}$. Ambiguous context size $k = 3$. For sample sizes $n$ see Appendix A.9.

| Method | Source of $\mathcal{P}$ | Max #desc. ↓ | ImageNet | ImageNetV2 | CUB200 | EuroSAT | Places365 | DTD | Flowers102 |
|---|---|---|---|---|---|---|---|---|---|
| CLIP | - | 1 | 61.87 | 54.74 | 51.69 | 40.92 | 39.01 | 43.09 | 62.81 |
| DCLIP | DCLIP | 12 | 62.22 | 54.84 | 52.55 | 47.33 | 40.01 | 41.86 | 62.17 |
| WaffleClip | WaffleClip | 30 | 63.31 | <u>55.92</u> | 52.38 | 44.31 | 40.56 | 43.16 | 66.27 |
| FuDD | FuDD | 1842 | **64.19** | **56.75** | 54.30 | 45.18 | 42.17 | 44.84 | 67.62 |
| *Ours* | DCLIP | 5 | 61.59 | 53.61 | 55.89 | <u>50.05</u> | 42.77 | **48.83** | 66.99 |
| *Ours* | Contrastive | 20 | 63.30 | 55.24 | 56.27 | **58.57** | **43.65** | <u>48.09</u> | <u>68.61</u> |
| *Ours* | FUDD | 25 | 61.86 | 53.05 | **56.62** | 48.42 | 42.76 | 48.03 | 68.47 |
| *Ours* | Contrastive | 50 | <u>63.51</u> | 55.41 | <u>56.45</u> | 44.46 | <u>43.62</u> | 47.66 | **69.51** |

Table 3: Image classification with classname included in the descriptions. Ambiguous context size $k = 3$. For sample sizes $n$ see Appendix A.9. An ablation of ambiguous context size $k$ can be found in Appendix A.5.

**Conventional Setup.** We evaluate our chosen descriptions in a conventional setup where classnames are included in all descriptions, as shown in Table 3. Our method performs well on datasets where a higher performance bump is observed with low $w_{cls}$, *i.e.* high relative description weights in Figure 3. This happens when the selected description pool provides generalizable, diverse, and discriminative descriptions for the datasets. Our method outperforms DCLIP assignments in the DCLIP pool and outdoes WaffleCLip and FuDD on datasets CUB200, EuroSAT, Places365, DTD, and Flowers102. This remains true for 4 of these datasets, even if only 5 descriptions per class are used. In contrast, WaffleClip (Roth et al. 2023) uses 30 text prompts per class, and FuDD (Esfandiarpoor and Bach 2024) uses an astonishing number of 1,842 descriptions per class.

On the other hand, for ImageNet and ImageNetV2, we can see a connection between suboptimal conventional performance and a much lower peak relative to baseline CLIP in the classname-free setting - indicating less distinctive power of our assignments. Mixed results in the conventional setup for ImageNet and ImageNetV2 imply it is challenging to find distinctive descriptions - at least within the currently used description pools $\mathcal{P}$. This difficulty may arise because random image contents, *e.g.*, background objects, distort the description assignments. Our algorithm experiences a performance boost when using a $\mathcal{P}$ obtained through contrastive prompting, offering a richer pool of descriptions.

Overall, the results from the class name-containing scenario suggest that the added semantics of the discovered descriptions enhance the performance—in addition to the class name ensembling used by other methods like WaffleClip, FuDD, and DCLIP.

**Performance in Classname-free Scenario with $w_{cls} = 0$.** We evaluate the performance under a classname-free scenario in Table 2 without any guiding classname information. In this case, random assignments don't achieve any reasonable classification accuracy; LLM-assigned descriptions provide minor guidance. With our selected descriptions, however, we have achieved decent performance across all datasets - significantly surpassing the LLM assignments. This further supports the idea that descriptions assigned by an LLM are not distinctive enough. Instead, feedback from the embedding space is needed for distinctive assignments. Higher distinctiveness also shows in Appendix A.6 where classname ensembling is prohibited via a maxing-aggregation.

## 5 Conclusion

This study demonstrates that *VLM Classification performance indeed benefits from LLM description semantics - if the descriptions are correctly selected*. To achieve this, we introduce a training-free method that assigns semantically meaningful descriptions based on feedback from the VLM embedding space. Our results indicate that these descriptions possess inherent discriminative power, as evidenced by evaluations conducted without classname ensembling in our proposed setup. Furthermore, incorporating these description assignments enhances performance in image classification tasks, both with and without classname ensembling. Additionally, our evaluation framework effectively distinguishes performance improvements arising from genuine semantic understanding from those resulting from ensemble effects. We hope that our findings will inspire future research on VLMs and contribute to the development of models with enhanced explainability.

## Acknowledgements

This project has been supported by the German Federal Ministry for Economic Affairs and Climate Action within the project "NXT GEN AI METHODS – Generative Methoden für Perzeption, Prädiktion und Planung", the German Research Foundation (DFG) project 421703927, Bayer AG, and the bidt project KLIMA-MEMES. The authors gratefully acknowledge the Gauss Center for Supercomputing for providing compute through the NIC on JUWELS at JSC and the HPC resources supplied by the Erlangen National High Performance Computing Center (NHR@FAU funded by DFG).

## References

Belinkov, Y.; and Bisk, Y. 2018. Synthetic and Natural Noise Both Break Neural Machine Translation. In *ICLR*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, J.; Yang, Z.; and Yang, D. 2020. MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification. In *ACL*.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing Textures in the Wild. In *CVPR*.

Cohen, J. M.; Rosenfeld, E.; and Kolter, J. Z. 2019. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Esfandiarpoor, R.; and Bach, S. H. 2024. Follow-Up Differential Descriptions: Language Models Resolve Ambiguities for Image Classification. *ICLR*.

Feng, S. Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; and Hovy, E. 2021. A survey of data augmentation approaches for NLP. *arXiv preprint arXiv:2105.03075*.

Feng, Z.; Bair, A.; and Kolter, J. Z. 2023. Text Descriptions are Compressive and Invariant Representations for Visual Learning. arXiv:2307.04317.

Hakimov, S.; and Schlangen, D. 2023. Images in Language Space: Exploring the Suitability of Large Language Models for Vision & Language Tasks. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics.

Hao, X.; Zhu, Y.; Appalaraju, S.; Zhang, A.; Zhang, W.; Li, B.; and Li, M. 2023. Mixgen: A new multi-modal data augmentation. In *WACV*.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2017. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. arXiv:2102.05918.

Jin, H.; Li, Z.; Tong, R.; and Lin, L. 2018. A deep 3D residual CNN for false-positive reduction in pulmonary nodule detection. *Medical physics*, 45(5): 2097–2107.

Kobayashi, S. 2018. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In *NAACL-HLT*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Menon, S.; and Vondrick, C. 2023. Visual Classification via Description from Large Language Models. *ICLR*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated Flower Classification over a Large Number of Classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Pratt, S.; Covert, I.; Liu, R.; and Farhadi, A. 2023. What does a platypus look like? Generating customized prompts for zero-shot image classification. arXiv:2209.03320.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, 5389–5400. PMLR.

Roth, K.; Kim, J. M.; Koepke, A.; Vinyals, O.; Schmid, C.; and Akata, Z. 2023. Waffling around for performance: Visual classification with random words and broad concepts. In *ICCV*, 15746–15757.

Şahin, G. G. 2022. To Augment or Not to Augment? A Comparative Study on Text Augmentation Techniques for Low-Resource NLP. *Computational Linguistics*.

Shtedritski, A.; Rupprecht, C.; and Vedaldi, A. 2023. What does clip know about a red circle? visual prompt engineering for vlms. In *ICCV*, 11987–11997.

Sun, L.; Xia, C.; Yin, W.; Liang, T.; Philip, S. Y.; and He, L. 2020. Mixup-Transformer: Dynamic Data Augmentation for NLP Tasks. In *Computational Linguistics*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology.

Yan, A.; Wang, Y.; Zhong, Y.; Dong, C.; He, Z.; Lu, Y.; Wang, W.; Shang, J.; and McAuley, J. 2023. Learning Concise and Descriptive Attributes for Visual Recognition. arXiv:2308.03685.

Yang, Y.; Panagopoulou, A.; Zhou, S.; Jin, D.; Callison-Burch, C.; and Yatskar, M. 2023. Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification. arXiv:2211.11158.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Zang, Y.; Yun, T.; Tan, H.; Bui, T.; and Sun, C. 2024. Pretrained Vision-Language Models Learn Discoverable Visual Concepts. arXiv:2404.12652.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhao, T.; Zhang, T.; Zhu, M.; Shen, H.; Lee, K.; Lu, X.; and Yin, J. 2022. An Explainable Toolbox for Evaluating Pre-trained Vision-Language Models. In Che, W.; and Shutova, E., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 30–37. Abu Dhabi, UAE: Association for Computational Linguistics.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million Image Database for Scene Recognition. *T-PAMI*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, 130(9): 2337–2348.