Optimization of Prompt Learning via Multi-Knowledge Representation for Vision-Language Models

Enming Zhang[1] Bingke Zhu[2] Yingying Chen[2]
Qinghai Miao[1*] Ming Tang[2] Jinqiao Wang[1,2]
[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2] Foundation Model Research Center, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
zhangenming23,miaoqh@mails.ucas.ac.cn
{bingke.zhu,yingying.chen,tangm,jqwang}@nlpr.ia.ac.cn

Abstract

Vision-Language Models (VLMs), such as CLIP, play a foundational role in various cross-modal applications. To fully leverage VLMs' potential in adapting to downstream tasks, context optimization methods like Prompt Tuning are essential. However, one key limitation is the lack of diversity in prompt templates, whether they are handcrafted or learned through additional modules. This limitation restricts the capabilities of pretrained VLMs and can result in incorrect predictions in downstream tasks. To address this challenge, we propose Context Optimization with Multi-Knowledge Representation (CoKnow), a framework that enhances Prompt Learning for VLMs with rich contextual knowledge. To facilitate CoKnow during inference, we trained lightweight semantic knowledge mappers, which are capable of generating Multi-Knowledge Representation for an input image without requiring additional priors. Experimentally, We conducted extensive experiments on 11 publicly available datasets, demonstrating that CoKnow outperforms a series of previous methods. We will make all resources open-source: https://github.com/EMZucas/CoKnow.

1. Introduction

By learning representations of millions of image-text pairs in a shared embedding space, large-scale pre-trained vision-language models (VLMs), *e.g.*, CLIP [22], have become powerful tool for a wide range of multimodal applications. The performance of VLMs is crucial for downstream tasks or applications that rely on VLMs as core components. However, plain VLMs may not perform optimally in some fundamental tasks, leading to unmet expectations. Figure 1(a) illustrates the classification task using CLIP on the CIFAR-10 [1] dataset. In this task, a specific template is used for text input, such as *a photo of a [CLASS]*, where *[CLASS]* is replaced with all categories. The text encoder generates the embedding for each category, while the image encoder produces the image embedding. These embeddings are then compared pairwise, with the category exhibiting the highest similarity to the image embedding being chosen as the final classification. For the case shown in Figure 1(a), CLIP incorrectly identifies an image of a bird as a frog, scoring 82% for this misclassification.

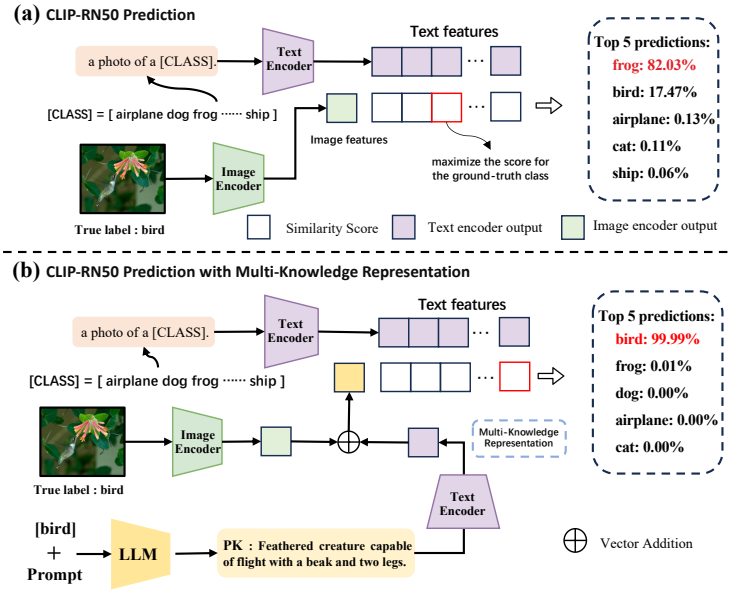

Figure 1. Comparison before and after introducing Multi-Knowledge Representation in zero-shot scenarios.

To enhance the performance of CLIP [22] in downstream tasks, many methods have been proposed, roughly catego-

*Corresponding author

rized into two main types: Fine Tuning and Prompt Learning. To avoid the high computational load of full parameter tuning, Fine Tuning methods typically freeze CLIP's parameters and introduce trainable additional modules to adapt to downstream tasks. Prompt Learning methods, on the other hand, attempt to learn better contextual templates to replace manually crafted prompts such as *a photo of a [CLASS]*. These methods effectively improve the performance of CLIP, but there is still much room for improvement in many tasks, especially in few-shot tasks in low-resource scenarios.

The improvement space for pretrained VLMs in downstream tasks can come from the potential they have yet to be elicited. One possible explanation is that a simple text prompt may not adequately represent the complexity of an image, which typically contains more information than its captions. In reality, a single object can be described in various ways, depending on the level of detail or perspective. While models like CLIP have been trained on a vast dataset of image-text pairs and have learned rich relationships between visual and textual content, they may struggle with simple text prompts that fail to capture the full semantic meaning of an image. To fully utilize the capabilities of CLIP, we propose to enhance the prompt context by incorporating knowledge from multiple perspectives at multiple abstraction levels, or in short Multi-Knowledge. In downstream tasks of image recognition, we define the global representation of a piece of text in VLMs as Multi-Knowledge Representation. Specifically, a natural language description consists of multiple words, each represented by a token in VLMs. Each word also represents different pieces of knowledge, and their encoded global representations constitute the Multi-Knowledge Representation.

The Multi-Knowledge introduced in this study comprise three types. The first one is visual knowledge (VK), which includes captions describing the image or its category. For the example in Figure 1, the visual knowledge for the given image could be *Feathers, beak, wings, two-legged, small body size.* Secondly, and of greater importance, we introduce non-visual knowledge (NVK) at a more abstract level beyond purely visual aspects. The NVK for the image in Figure 1 could be *Migratory creatures mastering flight and song.* Thirdly, we can combine multilevel descriptions, such as both VK and NVK, into a more comprehensive description called panoramic knowledge (PK). As shown in Figure 1(b), the PK could be *Feathered creature capable of flight with a beak and two legs.* With the introduction of PK, encoded and added to the image embedding, the prediction for the input image, previously predicted as *frog* in Figure 1(a), is now predicted as the correct category *bird* with 99.99% confidence. Furthermore, experimental results across the entire CIFAR-10 test set of 10,000 images show a 7.64% increase in prediction accuracy simply by introducing PK. *(The specific experimental details and results can be found in the appendix.)* This study demonstrates the essential role of Multi-Knowledge Representation in enhancing downstream tasks.

To leverage Multi-Knowledge Representation and the successful techniques of fine-tuning and soft prompt tuning for VLMs, we propose CoKnow, which consists of two learnable key modules, as shown in Figure 2. The first module is an optimizer for Prompt Learning guided by Multi-Knowledge representation, allowing for adaptive learning of prompt templates rich in domain knowledge for various downstream tasks. The second module is a lightweight semantic knowledge mapper, which, once trained, can directly generate the corresponding Multi-Knowledge Representation from images without additional input. This mapper is designed as a general module that can operate in a plug-and-play style with general VLMs, not limited to CLIP. Notably, CoKnow leverages the Large Language Model GPT-4 [2] as the knowledge generator, automatically producing Multi-Knowledge Representation using a set of simple prompt templates. Experimentally, We conducted experiments on 11 publicly available datasets spanning various downstream domains using CoKnow. The results, surpassing several previous few-shot methods, confirm that CoKnow is an effective approach for VLM applications. Our main contributions are as follows:

- To enhance VLMs performance in downstream tasks like classification, We propose Multi-Knowledge Representation, including image captions at visual level, non-visual descriptions at abstract level, as well as their combinations.
- We introduce the CoKnow framework, featuring a trainable prompt optimizer and lightweight visual-text semantic mappers, to support Multi-Knowledge Representation in tasks based on vision-language models.
- Effective and scalable methods are provided by us for generating knowledge across diverse categories. For 11 publicly available datasets, we construct 6603 distinct knowledge descriptions corresponding to 2201 object categories using GPT-4.

2. Related Work

Nowadays, significant advancements are being made in vision-language models [13, 23, 25, 31, 32], which create rich multimodal representations between natural language and vision, facilitating a wide array of downstream tasks. Among these models, CLIP [22] stands out as a prominent method. By undergoing self-supervised training on 400 million image-text pairs, CLIP learns joint representations for images and text, showcasing outstanding performance across various downstream visual tasks. Although pre-trained VLMs can represent image-text pairs, applying them effectively to downstream tasks still faces severasl

Table 1. Comparison of image recognition methods from different series for vision-language models. Extra Modules: Refers to incorporating other network modules for training. Soft Prompt: Refers to learning adaptable templates using learnable embedding vectors for downstream tasks. Image Rep: Represents the utilization of image representation. Fixed Template Rep: Represents the utilization of fixed templates from the downstream dataset, such as *a photo of a [classname]*. Multi-Knowledge Rep: Represents the utilization of multi-knowledge representation. ✓indicates that this part of the content is utilized in the method.

Methods	Extra Modules	Soft Prompt	Image Rep.	Fixed Template Rep.	Multi-Knowledge Rep.
CLIP-Adapter [8]	✓	-	✓	✓	-
Tip-Adapter [34]	✓	-	✓	✓	-
CLIP-A-self [18]	✓	-	✓	-	✓
CoOp [36]	-	✓	✓	-	-
CoCoOp [35]	✓	✓	✓	-	-
kgCoOp [30]	✓	✓	✓	✓	-
CoKnow(Ours)	✓	✓	✓	✓	✓

challenges. Recently, many studies have shown improved performance in downstream tasks, such as few-shot image recognition [18, 34, 36], by customizing CLIP through Fine Tuning and Prompt Learning.

2.1. Fine Tuning

In downstream tasks with few samples, full-finetuning [14, 19, 33] of CLIP often leads to overfitting issues, resulting in poor generalization performance on test data. Recently, introducing additional modules to CLIP and fine-tuning them has proven to be an effective method for adapting it to downstream image recognition tasks. CLIP adopts linear probing [4, 22], which involves introducing an additional linear layer for fine-tuning on downstream tasks. CALIP [10] invoked a parameter-free attention mechanism to promote interactions between visual and textual modalities, exploring essential cross-modal informational features. Its parametric iteration, CALIP-F [10], particularly excelled in few-shot testing environments, showcasing outstanding efficiency. WiSE-FT [28] proposed a weight-space ensemble approach, merging CLIP's pre-trained and fine-tuned weights to bolster robustness against atypical distributions. Houlsby and colleagues [12] introduced an adapter strategy that embeds learnable linear layers into transformer layers while immobilizing the network's backbone, providing a streamlined path for fine-tuning downstream tasks. Gao et al. [8] formulated the CLIP-Adapter, integrating feature adapters for nuanced fine-tuning within the visual or language branches. Zhang et al. Reference [34] launched the Tip-Adapter, which employs a key-value cache model conceived from a limited-sample training set, enhancing initialization for quicker fine-tuning convergence. Lin et al. [16] ventured into cross-modal few-shot learning by repurposing textual features as training samples during the fine-tuning phase, paving a novel pathway for utilizing textual data in training. These methods adapt CLIP for downstream tasks by introducing additional modules. Similarly, we trained the semantic knowledge mappers, which can automatically generate corresponding Multi-Knowledge Representation without the need for additional input, offering the advantage of plug-and-play functionality.

2.2. Prompt Learning

Another major challenge of applying large-scale pre-trained VLMs like CLIP to image recognition tasks is prompt engineering, which demands domain-specific expertise downstream and is extremely time-consuming, requiring substantial time investment for adjustments. To address this issue, CoOp [36] first introduced prompt learning into CLIP, achieving outstanding results in downstream few-shot image recognition tasks. Subsequent improvements have enabled a variety of tasks, including fine-grained object retrieval [27] and image-text classification [9]. Furthermore, CoCoOp [35] enhanced generalization by incorporating image features into each prompt embedding using lightweight networks. ProGrad [37] refined downstream task adaptation by correcting feedback gradients. Recently, KgCoOp [30] proposed constraining learnable prompt embeddings with common knowledge to enhance transfer generalization. These methods optimize learnable templates by leveraging image representations, and further, KgCoOp introduces hand-crafted template representations for optimizing learnable templates.

2.3. Multi-Knowledge Representation

Both existing methods, Prompt Learning and Model Fine-tuning, overlook the potential for incorporating a broader range of domain-specific knowledge representations corresponding to image representations, which can be articulated using natural language. We refer to this global semantic representation described in natural language as Multi-Knowledge Representation. Recently, CLIP-A-self [18] leverages GPT-4 to generate multiple visual description texts as input to the text encoder, and utilizes self-

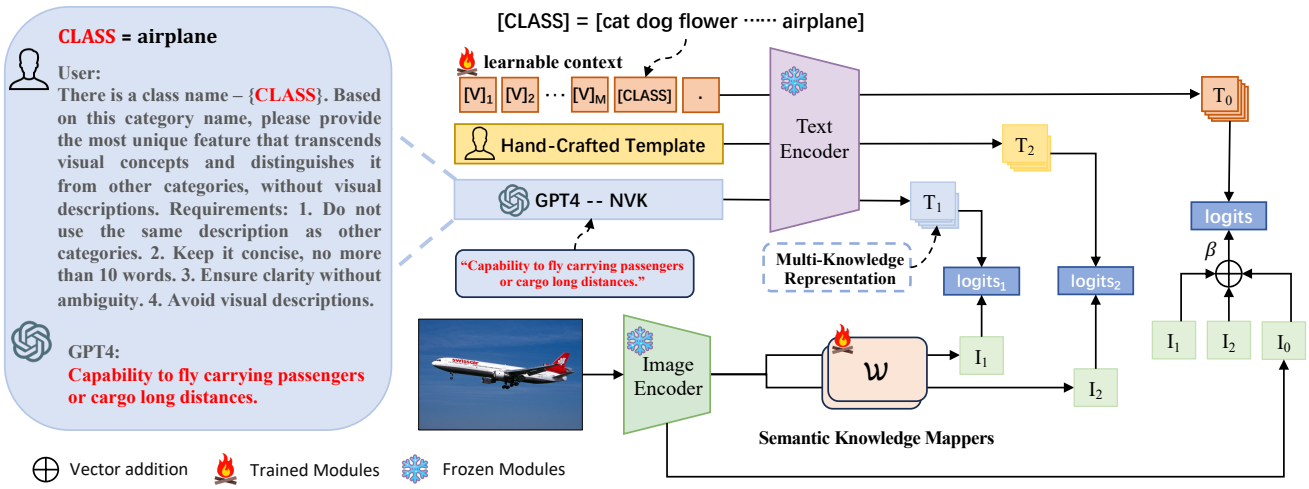

Figure 2. The overall architecture and training process of CoKnow. The figure illustrates the input of three types of templates into the text encoder: Learnable Context, Multi-Knowledge, and Hand-Crafted Template, when the category is *airplane*. The output of the image encoder undergoes contrastive loss calculation with the corresponding target template embeddings after passing through the semantic knowledge mappers. The original image representation and the mapped image representation are added with β weights before undergoing contrastive loss calculation with the final embedding representation of Learnable Contexts. Throughout the entire process, we only train lightweight semantic knowledge mappers to optimize context learning.

attention mechanism networks to aggregate them on each sentence, achieving good generalization effects. These visual descriptions are remarkably similar to the VK type within Multi-Knowledge. We first propose the use of Multi-Knowledge Representation to optimize learnable contextual templates, delivering domain knowledge information to downstream tasks through adaptive means, while reducing human involvement, thus making the process more convenient. And explored the effectiveness of other types of Multi-Knowledge, including VK, NVK, and PK.

In Table 1, we clearly delineate the distinctions between our method and various others. Existing approaches primarily focus on learning soft prompts and fine-tuning Extra Modules, lacking guidance from Multi-Knowledge Representation. Methods centered around domain knowledge require manual adjustments of this natural language aspect, lacking an adaptive template design approach. Our proposed method, CoKnow, addresses these shortcomings. For context template optimization, we not only leverage representations of images and fixed text but also introduce additional Multi-Knowledge Representation for context optimization, achieving outstanding results in experiments for the first time. Moreover, due to the presence of lightweight semantic knowledge mappers in CoKnow, no additional input is required during the inference stage. Prediction results can be obtained solely by inputting the image and the trained context template.

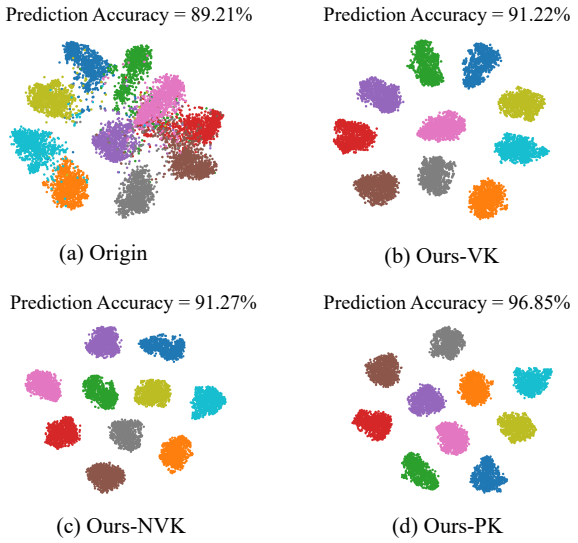

Figure 3. t-SNE visualization. Different colors represent different categories in the CIFAR-10 dataset. In the case of zero-shot, compared to the original CLIP, a clear division into 10 clusters can be observed.

3. Method

Our method is based on a pre-trained vision-language model CLIP [22], as shown in Figure 2. In terms of model architecture, we only introduce the lightweight semantic knowledge mappers, which can generate Multi-

Knowledge Representation without additional priors. The Multi-Knowledge Representation is important. We introduce three different types of Multi-Knowledge Representation, namely VK, NVK, and PK, respectively, on CIFAR-10 [1] dataset. By incorporating the Multi-Knowledge Representation with the corresponding original image representation, we obtain new representation vectors. *(Same as the approach in Figure 1(b).)* The results are visualized using t-SNE as depicted in Figure 3. *(For specific details and additional results, please refer to Appendix.)* It can be observed that introducing different types of Multi-Knowledge Representation allows for excellent partitioning of the new representations into 10 clusters. We have constructed a Prompt learning framework guided by Multi-Knowledge Representation. Next, we will introduce the technical details and principles of our approach.

3.1. Text Encoder Input

This section provides a detailed introduction to the construction of inputs for the text encoder during the training phase of CoKnow. The input consists of three main parts: Hand-Crafted Template, Multi-Knowledge, and Soft Prompt.

Hand-Crafted Template Since the introduction of CLIP, the method of inserting category names into Hand-Crafted Templates for image recognition has been widely adopted by subsequent researchers. In various downstream tasks, the approach of Hand-Crafted Templates has demonstrated superior performance. In CoKnow, we continue to leverage the Hand-Crafted Templates proposed by previous researchers, as shown in Figure 2. This type of hand-crafted template, such as *a photo of a [CLASS]*. Here, *[CLASS]* represents the insertable category name.

Multi-Knowledge We have defined three different types of Multi-Knowledge: Visual Knowledge (VK), Non-Visual Knowledge (NVK), and Panoramic Knowledge (PK: VK and NVK). The first one is visual knowledge (VK), which includes captions describing the image or its category. Secondly, and of greater importance, we introduce non-visual knowledge (NVK) at a more abstract level beyond purely visual aspects. Thirdly, we can combine multilevel descriptions, such as both VK and NVK, into a more comprehensive description called panoramic knowledge (PK). More instances of Multi-Knowledge can be found in Appendix. The differences between these different types of Multi-Knowledge lie in providing varied perspectives of knowledge description for the same image, beyond solely visual aspects. We generate the Multi-Knowledge using the large language model GPT-4 by constructing different prompts. As shown in Figure 2, the specific method for generating NVK is illustrated. When CLASS = *airplane* is input into the Prompt we constructed, we obtain the response from GPT-4: *Capability to fly carrying passengers or cargo long distances.* This represents the Multi-Knowledge described from a non-visual perspective, which is NVK. We utilize GPT-4 to construct Multi Knowledge for downstream tasks as input to the Text Encoder for training. In Appendix, we provide more detailed prompts for generating Multi-Knowledge.

Soft Prompt In CoKnow, we replace text with a set of continuous vectors as learnable templates. Let M be learnable vectors indexed as $\{v_1, v_2, ..., v_M\}$, where each vector corresponds to an embedding vector for a word. Let t_i represent the learnable prompt constructed for the i-th class, such that $t_i = \{v_1, v_2, ..., v_M, c_i\}$, where c_i represents the embedding vector for the class name of the i-th category. Similar to Hand-crafted Templates, each class name shares the same learnable prompt. As shown in Figure 2, the *learnable context* refers to Soft Prompt. We construct such Soft Prompts to learn effective templates adapted to downstream tasks.

3.2. Training of CoKnow

In this section, we have introduced the complete training framework of CoKnow. The entire training framework is illustrated in Figure 2, where we have only introduced lightweight semantic knowledge mappers as additional components to complete the entire architecture. In the text encoder, three types of templates are inputted: Learnable Context, Multi-Knowledge, and Hand-Crafted Template. At the output end of the text encoder, three sets of representation vectors are obtained: $T_0 = \{t_1^0, t_2^0, ..., t_k^0\}$, $T_1 = \{t_1^1, t_2^1, ..., t_k^1\}$, and $T_2 = \{t_1^2, t_2^2, ..., t_k^2\}$. Here, k represents the number of categories. After a single image is inputted into the image encoder, it is duplicated into two identical visual representation vectors. These vectors then undergo processing through two semantic knowledge mappers, resulting in output vectors I_1 and I_2. The logits are calculated between I_1 and T_1, as well as between I_2 and T_2. Simultaneously, the vectors I_0, I_1, and I_2 are added together using weight parameters β to obtain the vector I'. Finally, logits are calculated between the vector I' and T_0. The specific formulas are as follows:

$$\text{logits}_1 = \text{logit\_scale} \cdot I_1 T_1^T, \tag{1}$$

$$\text{logits}_2 = \text{logit\_scale} \cdot I_2 T_2^T, \tag{2}$$

$$I' = I_0 \cdot \beta + I_1 \cdot \frac{1-\beta}{2} + I_2 \cdot \frac{1-\beta}{2}, \tag{3}$$

$$\text{logits} = \text{logit\_scale} \cdot I' T_0^T, \tag{4}$$

where the logit_scale refers to a scaling factor. The logits_1,

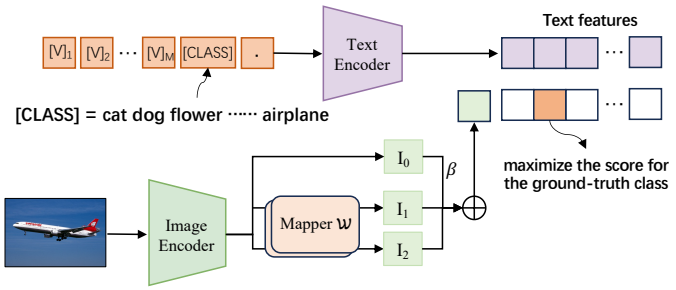

Figure 4. The inference stage of CoKnow. The prediction process of CoKnow involves automatically generating additional semantic knowledge representations when an image is inputted. These representations are then aggregated via β weights to produce a new vector for prediction. During inference, our method only requires the input image and the completed training contextual template, without any additional inputs needed.

logits_2, and logits are computed with the *label* using cross-entropy loss. Finally, the sum of the losses from these three logits is used to update the parameters. We only train the parameters of two lightweight semantic knowledge mappers w_1 and w_2 to optimize the context templates. Our semantic knowledge mappers adopt a three-layer fully connected neural network structure, consisting of an input layer, one hidden layer, and an output layer. We set the dimensionality of the hidden layer to one-fourth of the input dimension and add a ReLU activation function after the input layer.

3.3. Inference of CoKnow

In this section, we have introduced the complete inference framework of CoKnow. As illustrated in Figure 4, our method is very concise during the inference stage, requiring only the input of the image and the learned context template, without any additional input. Assuming there are K categories in total, where w_i represents the output of the text encoder when the category is i, x_0 represents the output of the image encoder for the given image I, and x_1 and x_2 represent the outputs of the semantic knowledge mappers W for the given image I, the probability of it belonging to category i is as follows:

$$x = \beta \cdot x_0 + (\frac{1-\beta}{2}) \cdot x_1 + (\frac{1-\beta}{2}) \cdot x_2, \tag{5}$$

$$p(y=i|I) = \frac{\exp(\text{sim}(w_i, x)/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(w_k, x)/\tau)}, \tag{6}$$

where $\text{sim}(w_i, x)$ denotes the cosine similarity between w_i and x, β represents the weight parameter,and τ is a temperature parameter.

4. Experiments

4.1. Experimental Settings

Datasets We conducted experiments on 11 publicly available datasets, encompassing a wide range of scenes and rich content, including ImageNet [6], Caltech [7], OxfordPets [21], Flowers [20], Food101 [3], Stanford Cars [15], FGVCAircraft [17], EuroSAT [11], UCF101 [26], DTD [5], and SUN397 [29]. In strict adherence to the few-shot evaluation protocol outlined in CoOp [36], we utilized 1, 2, 4, 8, and 16 shots for training, respectively, and validated the results on the complete test sets.

Implementation Details To ensure a fair comparison, we strictly adhered to the CoOp protocol [36], employing ResNet-50 as the backbone architecture for the CLIP image encoder. Additionally, we evaluated the performance using ViT-B/16 as the vision encoder. Our experiments were conducted with a batch size of 32, epoch sizes ranging from 10 to 400, and the β value set to 0.6. Learning rates of 0.001 and 0.002 were utilized, alongside the SGD optimizer and a cosine annealing strategy. The training process was executed on 6 NVIDIA GeForce RTX 3090 GPUs. We utilize three types of Multi-Knowledge through GPT4, including visual knowledge (VK), non-visual knowledge (NVK), and panoramic knowledge (PK: VK and NVK).

4.2. Performance

To assess the effectiveness of our approach, we conducted fair comparisons with a range of prior works, including CoOp [36], CoCoOp [35], Wise-FT [28], ProGrad [37], Tip-Adapter-F [34], Cross-Modal Wise-FT [16], and Linear-Probe CLIP [22].

Few-shot Learning In Table 2, we present the comparison of the average top-1 accuracy rates of CoKnow with CoOp across each dataset. This intuitively detailed display showcases the performance difference between CoKnow and prompt learning methods solely utilizing image representations, with CoOp serving as our baseline. Notably, in every accuracy metric, we have demonstrated superior performance. In Table 3, we have compared our approach with a greater variety of methods than before. Our method surpasses a range of previous approaches from 2 shots to 16 shots. It is worth noting that in the 1 shot scenario, CoKnow outperforms both CoOp and Wise-FT's 2 shots results. With 4 shots, CoKnow's results approach those of CoOp with 8 shots and surpass Wise-FT's results. In the 8 shots scenario, CoKnow's results are comparable to CoOp's and even surpass Wise-FT's results. These experimental results all demonstrate the effectiveness of CoKnow in learning from limited data.

Distribution Shift We further evaluated the robustness of CoKnow under out-of-distribution (OOD) conditions. We conducted prompt learning on ImageNet and tested on Ima-

Table 2. Per-dataset results on the ResNet-50 backbone. We bold the best result for each shot and each dataset. CoKnow consistently produces the best performance across all datasets.

Method	Shots	Dataset											
		Caltech [7]	ImageNet [6]	DTD [5]	EuroSAT [11]	Aircraft [17]	Food [3]	Flowers [20]	Pets [21]	Cars [15]	SUN397 [29]	UCF101 [26]	Average
Zero-Shot CLIP	0	86.29	58.18	42.32	37.56	17.28	77.31	66.14	85.77	55.61	58.52	61.46	58.77
CoOp	1	87.53	57.15	44.39	50.63	9.64	74.32	68.12	85.89	55.59	60.29	61.92	59.59
	2	87.93	57.81	45.15	61.50	18.68	72.49	77.51	82.64	58.28	59.48	64.09	62.32
	4	89.55	59.99	53.49	70.18	21.87	73.33	86.20	86.70	62.62	63.47	67.03	66.77
	8	90.21	61.56	59.97	76.73	26.13	71.82	91.18	85.32	68.43	65.52	71.94	69.89
	16	91.83	62.95	63.58	83.53	31.26	74.67	94.51	87.01	73.36	69.26	75.71	73.42
CoKnow	1	**89.30**	**59.17**	**47.93**	**58.67**	**18.97**	**76.59**	**75.83**	**86.94**	**57.66**	**61.38**	**64.69**	**63.38**
	2	**90.72**	**60.07**	**51.45**	**66.97**	**22.02**	**77.35**	**82.63**	**87.05**	**60.92**	**63.98**	**68.21**	**66.49**
	4	**90.90**	**61.00**	**58.23**	**74.60**	**25.67**	**77.70**	**88.67**	**88.17**	**65.67**	**66.03**	**71.93**	**69.87**
	8	**91.40**	**62.37**	**62.90**	**79.83**	**30.47**	**78.60**	**92.77**	**88.90**	**70.87**	**68.73**	**76.63**	**73.04**
	16	**92.83**	**63.83**	**67.20**	**84.77**	**37.57**	**78.83**	**95.33**	**89.13**	**76.67**	**71.07**	**79.77**	**76.09**

Table 3. Comparison with previous methods. We used ResNet50 as the backbone to compare the average top-1 accuracy across 11 test sets under the condition of few-shot learning with five different sample sizes. *Linear Probe* refers to the Linear-Probe CLIP method. The best results are highlighted in bold, and the second-best results are underlined.

Method	Ref.	Number of Shots				
		1	2	4	8	16
Linear Probe	–	36.67	47.61	57.19	64.98	71.10
CoOp	IJCV'22	59.59	62.32	66.77	69.89	73.42
Wise-FT	CVPR'22	59.09	61.80	65.29	68.43	71.64
ProGrad	ICCV'23	62.61	64.90	68.45	71.41	73.96
Tip-Adapter-F	ECCV'22	63.25	65.93	68.98	<u>72.15</u>	<u>75.10</u>
Cross-Modal Wise-FT	CVPR'23	**63.76**	<u>66.40</u>	68.95	71.73	74.08
CoKnow	–	<u>63.38</u>	**66.49**	**69.87**	**73.04**	**76.09**

Table 4. Comparison on Distribution Shift. In the context of distribution shift, we have conducted comparisons with previous works. We employed ViT-B/16 for training on the ImageNet dataset, followed by zero-shot testing on the ImageNet-V2 datasets.

Method	ImageNet	ImageNet-V2
Zero-Shot CLIP	66.7	60.8
Linear Probe	65.9	56.3
CoOp[36]	71.7	64.6
CoCoOp[35]	71.0	64.1
KgCoOp[30]	71.2	64.1
CoKnow	**72.07**	**64.7**

geNetV2 [24]. The results are presented in Table 4. The experimental results indicate that the method effectively generalizes from ImageNet to out-of-distribution datasets, showcasing its potential in handling distribution shifts.

4.3. Ablation Study

Different Multi-Knowledge. We utilize VK, NVK, and PK, three different types of knowledge descriptions, to further explore the impact of different Multi-Knowledge on prompt learning results under the condition of inputting the same Hand-crafted Template. Setting β at 0.8 and optimizing the context length to 16, we detailed the average effects of few-shot learning across 11 datasets in Figure 5. Overall, descriptions of varying Multi-Knowledge were able to further optimize prompt learning, achieving commendable levels of performance. The overall prediction accuracy of PK is the best, indicating that a global representation incorporating knowledge from multiple perspectives is more effective.

Different β weight values. We investigated the impact of different β weights on the model's outcomes by setting the β values to 0.4, 0.6, and 0.8, respectively, with PK as the input and the optimization context length set to 16. Figure 6 presents a comparison of the average predictions across 11 datasets under the condition of 16 shots. Simultaneously, the effects of previous methods are directly demonstrated in the graph. Different values of β have a certain impact on the results, but they all outperform a series of previous excellent methods. Our experiments offer guidance for adjusting the β value.

Ablation Study of Mapper. The experiment explores the effectiveness of using a semantic knowledge mapper to map the image representations of the original CLIP (denoted as I_0 in Figure 2), as well as investigates the importance of the original CLIP's image representations for Prompt Learning. The β value is set to 0.6, the input is set to PK, and the optimized context length is set to 16. We incorporate a semantic knowledge mapper before I_0 in Fig-

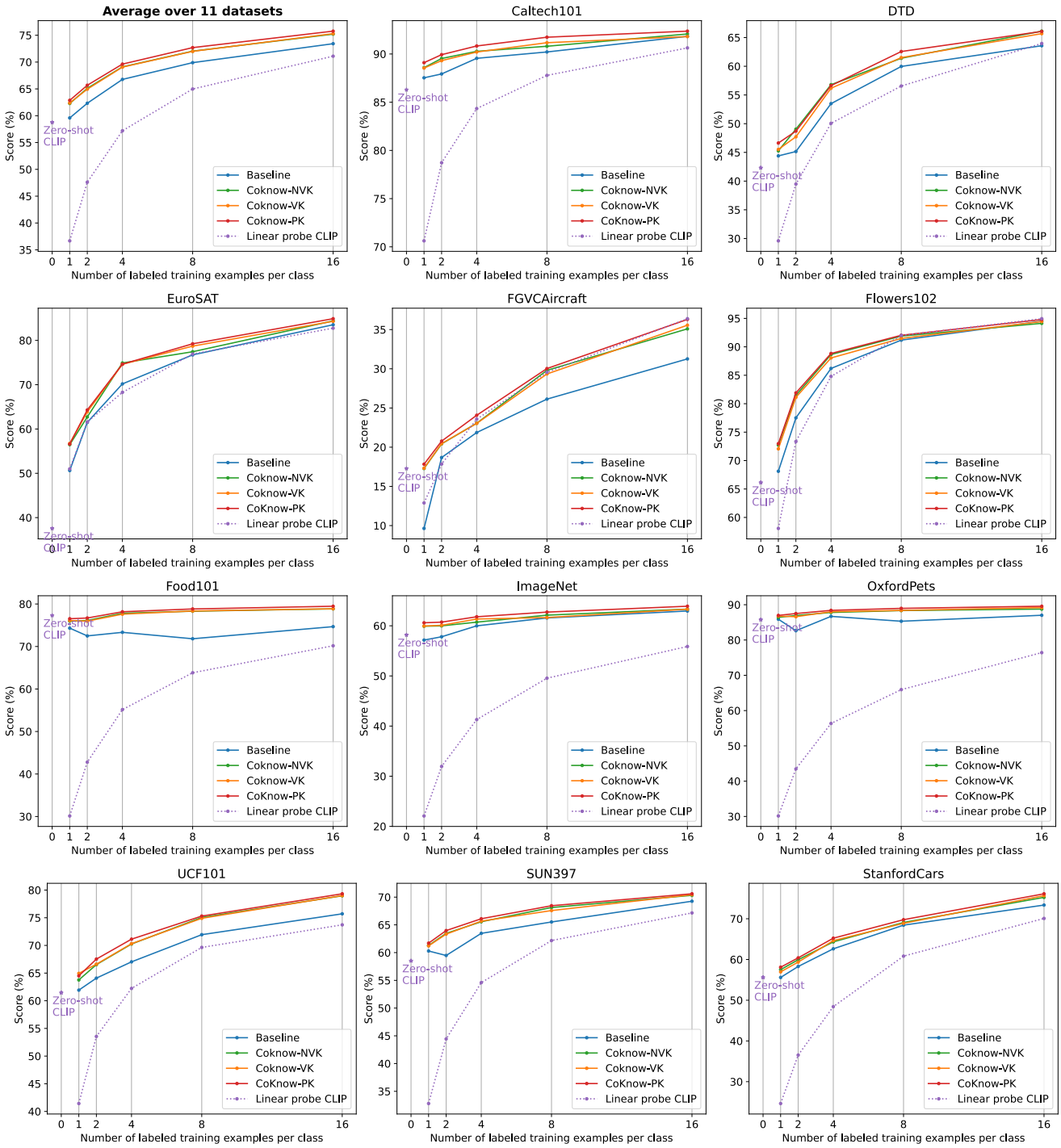

Figure 5. Ablation Study on Inputting Different Multi-Knowledge. At the text input end, apart from the Hand-crafted Template, we input different Multi-Knowledge, including NVK, VK, and PK, respectively, as target Multi-Knowledge Representation for learning. *The Baseline represents direct prompt learning via CoOp.*

ure 2, and the new structure is referred to as *CoKnow-I*. As shown in Table 5, it can be observed that when the training sample size is 1 shot, CoKnow-I achieves very low Top-1 accuracy across 11 datasets. When the training sample size is increased to 16 shots, the Top-1 accuracy across these 11 datasets remains lower than the predicted results of Co-

Table 5. Ablation Study of Mapper. CoKnow-I represents the re-mapping of the original CLIP's image representations in CoKnow using a semantic knowledge mapper.

Method	Shots	Dataset											
		Caltech [7]	ImageNet [6]	DTD [5]	EuroSAT [11]	Aircraft [17]	Food [3]	Flowers [20]	Pets [21]	Cars [15]	SUN397 [29]	UCF101 [26]	Average
CoKnow-I	1	47.23	27.40	27.67	53.07	13.13	13.33	56.27	10.13	32.13	26.03	22.43	29.89
	16	91.13	60.52	63.77	83.40	35.13	68.97	94.10	78.3	72.57	65.20	74.70	71.10
CoKnow	1	**89.30**	**59.17**	**47.93**	**58.67**	**18.97**	**76.59**	**75.83**	**86.94**	**57.66**	**61.38**	**64.69**	**63.38**
	16	**92.83**	**63.83**	**67.20**	**84.77**	**37.57**	**78.83**	**95.33**	**89.13**	**76.67**	**71.07**	**79.77**	**76.09**

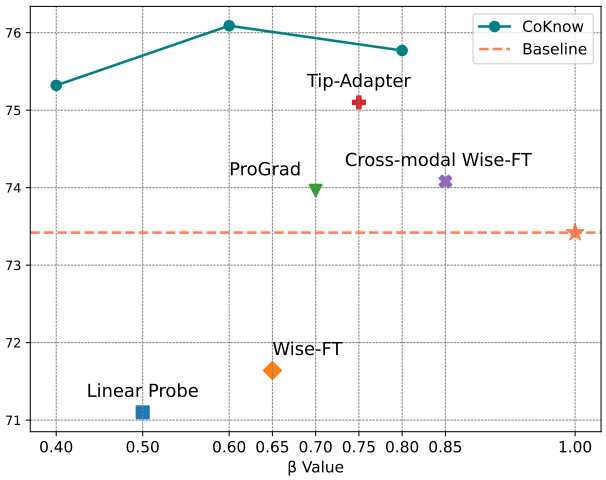

Figure 6. Average prediction results for different β values. The accuracy of a series of previous methods is directly marked on the graph.

Table 6. Ablation Study on Different Classname Positions. *End* represents the category name located at the end of the learnable vectors, while *Middle* represents the category name located in the middle of the learnable vectors.

Position of Classnames		Number of Shots				
End	Middle	1	2	4	8	16
✓		**63.31**	**66.00**	**69.87**	73.04	**76.09**
	✓	61.91	65.63	69.58	**73.11**	76.01

Table 7. Average Prediction Results for Different Context Lengths. This represents the lengths of learnable vectors of different sizes.

Context Length	Accuracy(%)
8	76.02
16	**76.09**
32	75.96

Know. The experimental results indicate that the image representations of the original CLIP have a significant impact on Prompt Learning, especially evident when the training sample size is 1-shot. Additionally, it is not advisable to use a semantic knowledge mapper to map the image representations of the original CLIP.

Position of Classname Varies. In the context of optimization, the position of the *classname* can be either at the end or in the middle [36]. Setting $\beta = 0.6$, optimizing the context length to 16 and PK as the input. We explored the impact of different *classname* positions on experimental results across 11 datasets. The outcomes are presented in Table 6. Overall performance is better when the *classname* position is at the end rather than in the middle.

Different Context Lengths. We set the context lengths to 8, 16, and 32, with PK as the input and the β value set to 0.6. We evaluated the mean comparison after training with 16 shots on 11 datasets to explore the impact of different context lengths on the results. The results are shown in Table 7. Our method demonstrates robust performance across varying context lengths, with minimal perturbation to the outcomes attributable to changes in context length.

5. CONCLUSION

In this paper, we have gained insights into the importance of Multi-Knowledge Representation for VLMs. Existing technologies have overlooked the method of context template learning, which incorporates domain knowledge. Such oversight will restrict VLM's ability to learn Multi-Knowledge Representation during the pre-training phase, further limiting its performance in downstream tasks. To address this issue, we propose Context Optimization with Multi-Knowledge Representation (CoKnow). We trained lightweight semantic knowledge mappers, capable of generating Multi-Knowledge Representation without requiring additional inputs, while also constructing a knowledge-guided Prompt learning framework. We extensively evaluated it on multiple benchmark tests, and the results indicate that the proposed CoKnow is an effective method for prompt learning.

References

[1] Yehya Abouelnaga, Ola S Ali, Hager Rady, and Mohamed Moustafa. Cifar-10: Knn-based ensemble of classifiers. In *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1192–1195. IEEE, 2016. 1, 5

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014. 6, 7, 9

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 6, 7, 9

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6, 7, 9

[7] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 6, 7, 9

[8] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 3

[9] Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. Texts as images in prompt tuning for multi-label image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2808–2817, 2023. 3

[10] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 746–754, 2023. 3

[11] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 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*, 12(7):2217–2226, 2019. 6, 7, 9

[12] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 3

[13] 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

[14] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 3

[15] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 6, 7, 9

[16] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19325–19337, 2023. 3, 6

[17] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6, 7, 9

[18] Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E O'Connor. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 262–271, 2023. 3

[19] Xiaofeng Mao, Yufeng Chen, Xiaojun Jia, Rong Zhang, Hui Xue, and Zhao Li. Context-aware robust fine-tuning. *International Journal of Computer Vision*, pages 1–16, 2023. 3

[20] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 6, 7, 9

[21] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6, 7, 9

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 6

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[24] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to im-

agenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 7

[25] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 2

[26] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6, 7, 9

[27] Shijie Wang, Jianlong Chang, Zhihui Wang, Haojie Li, Wanli Ouyang, and Qi Tian. Fine-grained retrieval prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2644–2652, 2023. 3

[28] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022. 3, 6

[29] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 6, 7, 9

[30] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023. 3, 7

[31] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[32] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 2

[33] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022. 3

[34] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 3, 6

[35] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 3, 6, 7

[36] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3, 6, 7, 9

[37] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023. 3, 6

A. Multi-Knowledge Generation

In this section, we offer detailed methodologies and concrete exemplars of Multi-Knowledge generation.

This paper investigates efficacious methodologies for Multi-Knowledge Generation across 14 datasets encompassing diverse fields. *(Table 8 provides a summary of these 14 datasets.)* The approach entails furnishing GPT-4 with scalable prompts to elicit its responses, thereby facilitating the acquisition of Multi-Knowledge. In Table 9, we furnish the prompts to generate various types of Multi-Knowledge across the 14 datasets, alongside the Hand-crafted Templates employed for these datasets.

Table 8. Datasets statistics.

Dataset	Classes	Train	Val	Test
ImageNet	1,000	1.28M	N/A	50,000
Caltech101	100	4,128	1,649	2,465
OxfordPets	37	2,944	736	3,669
StanfordCars	196	6,509	1,635	8,041
Flowers102	102	4,093	1,633	2,463
Food101	101	50,500	20,200	30,300
FGVCAircraft	100	3,334	3,333	3,333
SUN397	397	15,880	3,970	19,850
DTD	47	2,820	1,128	1,692
EuroSAT	10	13,500	5,400	8,100
UCF101	101	7,639	1,898	3,783
CIFAR-10	10	50,000	N/A	10,000
CIFAR-100	10	50,000	N/A	10,000
STL-10	10	5,000	N/A	8,000

In Figure 7, we present a specific instance to generate PK utilizing GPT-4. Upon providing a category *face* to GPT-4 as input, the comprehensive prompt yields the following response: *Area with eyes, nose, mouth; center of emotional expression.*

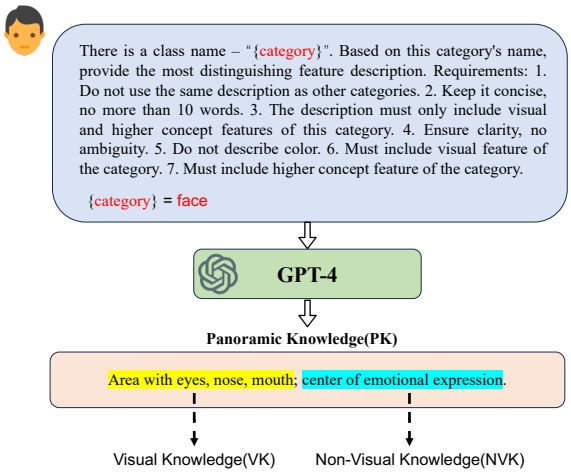

Figure 7. **The specific process of generating Multi-Knowledge.**

B. Instances of Multi-Knowledge

In this section, we provide detailed instances of Visual Knowledge (VK), Non-Visual Knowledge (NVK), and Panoramic Knowledge (PK). As shown in Figure 8, we randomly select multiple images and provide Multi-Knowledge generated based on their class names.

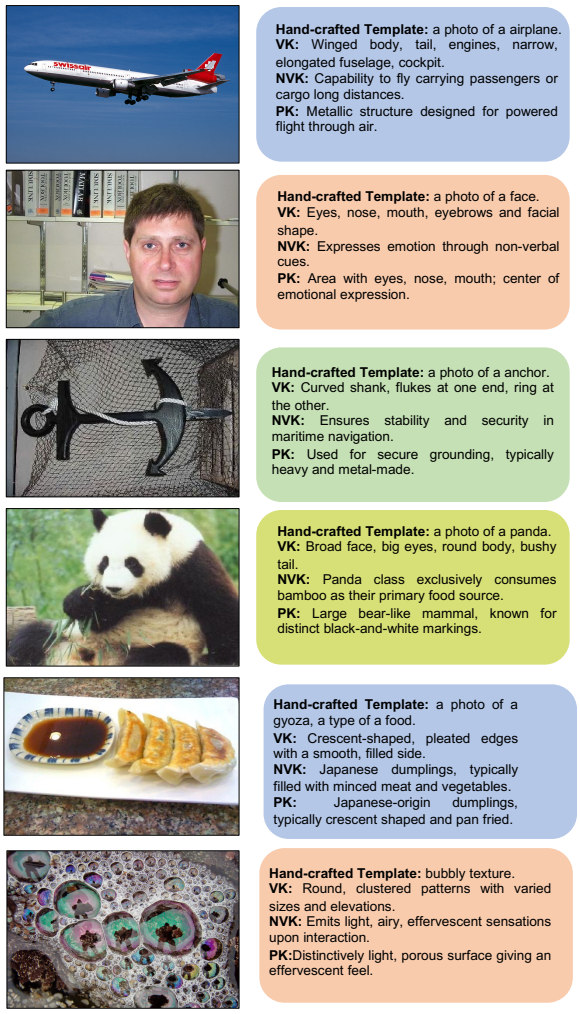

Figure 8. Instances of Multi-Knowledge.

C. Experiments of zero-shot prediction with Multi-knowledge

To explore the significance of Multi-Knowledge in pre-trained Vision-Language Models (VLMs), we conducted zero-shot experiments on the CIFAR-10, CIFAR-100, and STL10 datasets. We utilize Vision Transformer (ViT) as the visual encoding backbone for CLIP and perform experiments on the CIFAR-10 test set, comprising 10,000 images. Our prediction process is illustrated in Figure 9 *(which is identical to the method described in Figure 1 of the* ***main***

text). In Figure 9 (b), we use the PK, VK, and NVK generated by GPT-4 as inputs to the text encoder, to obtain the Multi-Knowledge Representation corresponding to the image.

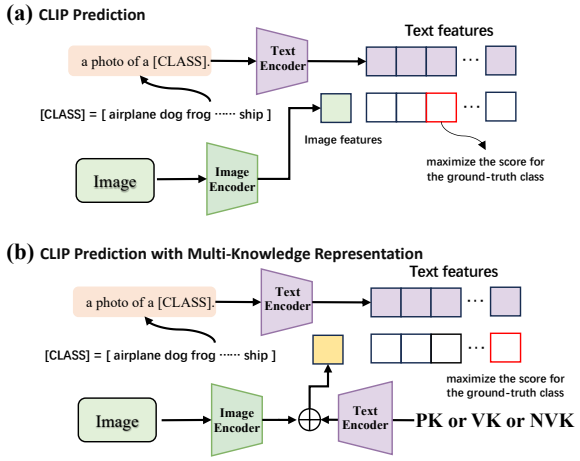

Figure 9. The prediction process in CLIP before and after the introduction of Multi-Knowledge.

In Figure 10, we present the zero-shot prediction with the incorporation of various types of Multi-Knowledge. Notably, PK yields the most substantial enhancement in accuracy. These experimental findings underscore the significance of Multi-Knowledge. As illustrated in Figure 11, we conduct a t-SNE visualization of the combined vector results from the original image representation and the Multi-Knowledge Representation. It can be observed that on various datasets, the representations are clearly segmented into multiple clusters, which correspond to the number of categories.

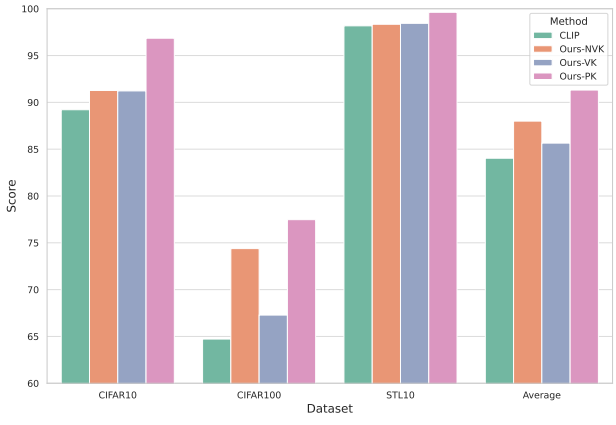

Figure 10. The prediction results with the introduction of different Multi-Knowledge. Among these results, PK achieves the best performance.

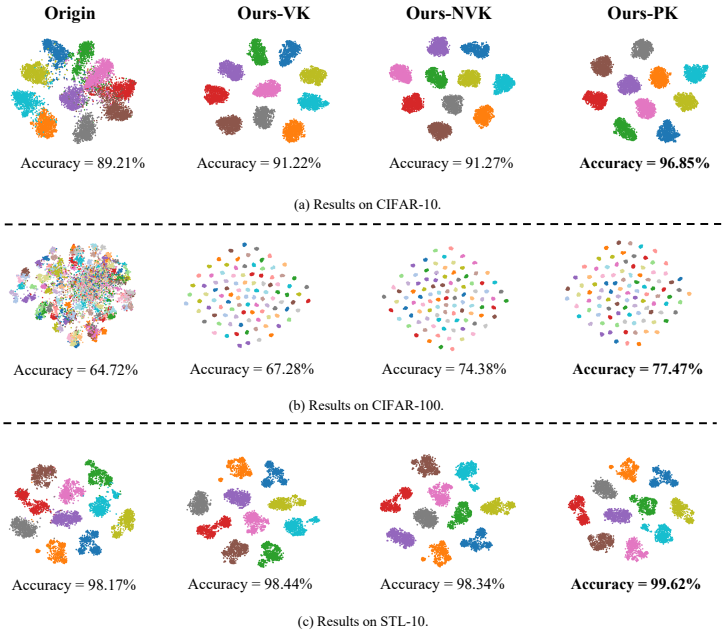

Figure 11. t-SNE visualization of the combined vector results from the original image representation and the Multi-Knowledge Representation. Different colors represent different categories.

Table 9. The prompts given to GPT-4 for Multi-Knowledge Generation.

Dataset	**VK**	**NVK**	**PK**	**Hand-crafted Template**
Imagenet	Based on the category name, please provide the most distinctive visual features unique to this category. Requirements: 1. Do not use the same appearance description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid mentioning colors.	Based on this category name, please provide the most unique feature that transcends visual concepts and distinguishes it from other categories, without visual descriptions. Requirements: 1. Do not use the same description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid visual descriptions.	Based on this category's name, provide the most distinguishing feature description. Requirements: 1. Do not use the same description as other categories. 2. Keep it concise, no more than 10 words. 3. The description must only include visual and higher concept features of this category. 4. Ensure clarity, no ambiguity. 5. Do not describe color. 6. Must include visual feature of the category. 7. Must include higher concept feature of the category.	a photo of a [CLASS].
Eurosat	This is the category name of images captured by satellites. Please describe the most distinctive visual feature unique to this category. Requirements: 1. Do not use the same appearance description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid mentioning colors.	This is the category name of images captured by satellites. Please provide the most unique feature that transcends visual concepts and distinguishes it from other categories, without visual descriptions. Requirements: 1. Do not use the same description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid visual descriptions.	This is the category name of images captured by satellites. Please provide the most distinguishing feature that transcends visual concepts from other categories. Requirements: 1. Do not use the same description as the satellite image category. 2. Keep it concise, no more than 10 words. 3. The description must only include visual and higher concept features of this category. 4. Ensure clarity, no ambiguity. 5. Do not describe color. 6. Must include the visual feature of this category. 7. Must include higher concept feature of this category.	a centered satellite photo of [CLASS].
Caltech101	Based on the category name, please provide the most distinctive visual features unique to this category. Requirements: 1. Do not use the same appearance description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid mentioning colors.	Based on this category name, please provide the most unique feature that transcends visual concepts and distinguishes it from other categories, without visual descriptions. Requirements: 1. Do not use the same description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid visual descriptions.	Based on this category's name, provide the most distinguishing feature description. Requirements: 1. Do not use the same description as other categories. 2. Keep it concise, no more than 10 words. 3. The description must only include visual and higher concept features of this category. 4. Ensure clarity, no ambiguity. 5. Do not describe color. 6. Must include visual feature of the category. 7. Must include higher concept feature of the category.	a photo of a [CLASS].

Continued on next page

Table 9 – continued from previous page

Dataset	VK Description	NVK Description	PK Description	Hand-crafted Desc.
FGVCaircraft	This is the category name of a type of aircraft model. Please provide a description of the most distinctive visual features unique to this category. Requirements: 1. Do not use the same appearance description as other aircraft models. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid mentioning colors.	This is the category name of a type of aircraft model. Please provide the most unique feature that transcends visual concepts and distinguishes it from other categories, without visual descriptions. Requirements: 1. Do not use the same description as other aircraft models. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid visual descriptions.	This is a category of aircraft models. Please provide the most distinguishing feature from other aircraft model categories. Requirements: 1. Do not use the same description as other aircraft model categories. 2. Keep it concise, no more than 10 words. 3. The description must only include visual and higher concept features of this category. 4. Ensure clarity, no ambiguity. 5. Do not describe color. 6. Must include the visual feature of this category. 7. Must include higher concept feature of this category.	a photo of a [CLASS],a type of aircraft.
Sun397	This is the name of a real-life scene category. Please provide a description of the most distinctive visual features unique to this category. Requirements: 1. Do not use the same appearance description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid mentioning colors.	This is the name of a real-life scene category. Please provide the most distinctive feature that transcends visual concepts and distinguishes it from other categories, without visual descriptions. Requirements: 1. Do not use the same description as other real-life scene categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid visual descriptions.	This is the category name of a type of real-life scene. Please provide a description of the most distinctive feature that sets it apart from other scene categories. Requirements: 1. Do not use the same description as other scene categories. 2. Keep it concise, no more than 10 words. 3. The description must only include visual and higher conceptual features of this category. 4. Ensure clarity without ambiguity. 5. Avoid mentioning colors. 6. Must include visual features of this category. 7. Must include higher conceptual features of this category.	a photo of a [CLASS].
Oxford pets	This is the name of an animal species. Please provide a description of the most distinctive visual features unique to this category. Requirements: 1. Do not use the same appearance description as other species. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid mentioning colors.	This is the name of a pet breed. Please provide the most distinctive feature that transcends visual concepts and distinguishes it from other breeds, without visual descriptions. Requirements: 1. Do not use the same description as other pet breeds. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid visual descriptions.	This is the category name of a type of pet breed. Please provide a description of the most distinctive feature that sets it apart from other pet breeds. Requirements: 1. Do not use the same description as other pet breed categories. 2. Keep it concise, no more than 10 words. 3. The description must only include visual and higher conceptual features of this category. 4. Ensure clarity without ambiguity. 5. Avoid mentioning colors. 6. Must include visual features of this category. 7. Must include higher conceptual features of this category.	a photo of a [CLASS], a type of pet.

Continued on next page

Table 9 – continued from previous page

Dataset	VK Description	NVK Description	PK Description	Hand-crafted Desc.
Stanford cars	This is the name of a category of cars. Please provide a description of the most distinctive visual features unique to this category. Requirements: 1. Do not use the same appearance description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid mentioning colors.	This is the name of a category of car brands. Please provide the most distinctive feature that transcends visual concepts and distinguishes it from other categories, without visual descriptions. Requirements: 1. Do not use the same description as other car brand categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid visual descriptions.	This is the category name of a type of car. Please provide a description of the most distinctive feature that sets it apart from other car categories. Requirements: 1. Do not use the same description as other car categories. 2. Keep it concise, no more than 10 words. 3. The description must only include visual and higher conceptual features of this category. 4. Ensure clarity without ambiguity. 5. Avoid mentioning colors. 6. Must include visual features of this category. 7. Must include higher conceptual features of this category.	a photo of a [CLASS].
Cifar10	Based on the category name, please provide the most distinctive visual features unique to this category. Requirements: 1. Do not use the same appearance description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid mentioning colors.	Based on this category name, please provide the most unique feature that transcends visual concepts and distinguishes it from other categories, without visual descriptions. Requirements: 1. Do not use the same description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid visual descriptions.	Based on this category's name, provide the most distinguishing feature description. Requirements: 1. Do not use the same description as other categories. 2. Keep it concise, no more than 10 words. 3. The description must only include visual and higher concept features of this category. 4. Ensure clarity, no ambiguity. 5. Do not describe color. 6. Must include visual feature of the category. 7. Must include higher concept feature of the category.	a photo of a [CLASS].
DTD	This is the name of a texture category. Please provide a description of the most distinctive visual features unique to this category. Requirements: 1. Do not use the same appearance description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid mentioning colors.	This is the name of a texture category. Please provide the most unique feature that transcends visual concepts and distinguishes it from other categories, without visual descriptions. Requirements: 1. Do not use the same description as other texture categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid visual descriptions.	This is the category name of a type of texture. Please provide a description of the most distinctive feature that sets it apart from other texture categories. Requirements: 1. Do not use the same description as other texture categories. 2. Keep it concise, no more than 10 words. 3. The description must only include visual and higher conceptual features of this category. 4. Ensure clarity without ambiguity. 5. Avoid mentioning colors. 6. Must include visual features of this category. 7. Must include higher conceptual features of this category.	[CLASS] texture.

Continued on next page

Table 9 – continued from previous page

Dataset	VK Description	NVK Description	PK Description	Hand-crafted Desc.
Food101	This is a category of food. Please provide the most distinctive visual feature of this category. Requirements: 1. Do not use the same appearance description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid mentioning colors.	This is the category name of a type of food. Please provide the most distinctive feature that transcends visual concepts and distinguishes it from other categories, without visual descriptions. Requirements: 1. Do not use the same description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid visual descriptions.	This is the name of a category of food. Please provide the most distinctive feature that sets it apart from other food categories. Requirements: 1. Do not use the same description as other food categories. 2. Keep it concise, no more than 10 words. 3. The description must only include visual and higher conceptual features of this category. 4. Ensure clarity without ambiguity. 5. Do not mention colors. 6. Must include visual features of this category. 7. Must include higher conceptual features of this category.	a photo of [CLASS], a type of food.
UCF101	This is the name of a character action behavior. Please provide a description of the most distinctive visual features unique to this action. Requirements: 1. Do not use the same appearance description as other actions. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid mentioning colors.	This is the name of a category of human behavior actions. Please provide the most distinctive feature that transcends visual concepts and distinguishes it from other categories, without visual descriptions. Requirements: 1. Do not use the same description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid visual descriptions.	This is the name of a human behavior action. Please provide a description of the most distinctive feature that sets it apart from other action behavior categories. Requirements: 1. Do not use the same description as other behavior categories. 2. Keep it concise, no more than 10 words. 3. The description must only include visual and higher conceptual features of this category. 4. Ensure clarity without ambiguity. 5. Avoid mentioning colors. 6. Must include visual features of this category. 7. Must include higher conceptual features of this category.	a photo of a person doing [CLASS].
Oxford flowers	This is the name of a category of flowers. Please provide a description of the most distinctive visual features unique to this category. Requirements: 1. Do not use the same appearance description as other flower categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid mentioning colors.	This is the name of a category of flowers. Please provide the most distinctive feature that transcends visual concepts and distinguishes it from other categories, without visual descriptions. Requirements: 1. Do not use the same description as other flower categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid visual descriptions.	This is the name of a category of flowers. Please provide a description of the most distinctive feature that sets it apart from other flower categories. Requirements: 1. Do not use the same description as other flower categories. 2. Keep it concise, no more than 10 words. 3. The description must only include visual and higher conceptual features of this category. 4. Ensure clarity without ambiguity. 5. Avoid mentioning colors. 6. Must include visual features of this category. 7. Must include higher conceptual features of this category.	a photo of a [CLASS], a type of flower.

Continued on next page

Table 9 – continued from previous page

Dataset	VK Description	NVK Description	PK Description	Hand-crafted Desc.
Cifar100	Based on the category name, please provide the most distinctive visual features unique to this category. Requirements: 1. Do not use the same appearance description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid mentioning colors.	Based on this category name, please provide the most unique feature that transcends visual concepts and distinguishes it from other categories, without visual descriptions. Requirements: 1. Do not use the same description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid visual descriptions.	Based on this category's name, provide the most distinguishing feature description. Requirements: 1. Do not use the same description as other categories. 2. Keep it concise, no more than 10 words. 3. The description must only include visual and higher concept features of this category. 4. Ensure clarity, no ambiguity. 5. Do not describe color. 6. Must include visual feature of the category. 7. Must include higher concept feature of the category.	a photo of a [CLASS].
Stl10	Based on the category name, please provide the most distinctive visual features unique to this category. Requirements: 1. Do not use the same appearance description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid mentioning colors.	Based on this category name, please provide the most unique feature that transcends visual concepts and distinguishes it from other categories, without visual descriptions. Requirements: 1. Do not use the same description as other categories. 2. Keep it concise, no more than 10 words. 3. Ensure clarity without ambiguity. 4. Avoid visual descriptions.	Based on this category's name, provide the most distinguishing feature description. Requirements: 1. Do not use the same description as other categories. 2. Keep it concise, no more than 10 words. 3. The description must only include visual and higher concept features of this category. 4. Ensure clarity, no ambiguity. 5. Do not describe color. 6. Must include visual feature of the category. 7. Must include higher concept feature of the category.	a photo of a [CLASS].