V2C-CBM: Building Concept Bottlenecks with Vision-to-Concept Tokenizer

Hangzhou He[1,2,3,4], **Lei Zhu**[1,2,3,4], **Xinliang Zhang**[1,2,3,4], **Shuang Zeng**[1,2,3,4], **Qian Chen**[1,2,3,4], **Yanye Lu**[1,2,3,4*]

[1]Institute of Medical Technology, Peking University, Beijing, China
[2]Department of Biomedical Engineering, Peking University, Beijing, China
[3]National Biomedical Imaging Center, Peking University, Beijing, China
[4]Institute of Biomedical Engineering, Peking University Shenzhen Graduate School, Shenzhen, China
{zhuang, zhulei, chen_qian}@stu.pku.edu.cn, zhangxinliang_mit@163.com, {stevezs, yanye.lu}@pku.edu.cn

Abstract

Concept Bottleneck Models (CBMs) offer inherent interpretability by initially translating images into human-comprehensible concepts, followed by a linear combination of these concepts for classification. However, the annotation of concepts for visual recognition tasks requires extensive expert knowledge and labor, constraining the broad adoption of CBMs. Recent approaches have leveraged the knowledge of large language models to construct concept bottlenecks, with multimodal models like CLIP subsequently mapping image features into the concept feature space for classification. Despite this, the concepts produced by language models can be verbose and may introduce non-visual attributes, which hurts accuracy and interpretability. In this study, we investigate to avoid these issues by constructing CBMs directly from multimodal models. To this end, we adopt common words as base concept vocabulary and leverage auxiliary unlabeled images to construct a Vision-to-Concept (V2C) tokenizer that can explicitly quantize images into their most relevant visual concepts, thus creating a vision-oriented concept bottleneck tightly coupled with the multimodal model. This leads to our V2C-CBM which is training efficient and interpretable with high accuracy. Our V2C-CBM has matched or outperformed LLM-supervised CBMs on various visual classification benchmarks, validating the efficacy of our approach.

Code — https://github.com/riverback/V2C-CBM

Introduction

With the increasing adoption of deep learning-based methods in high-risk and sensitive fields such as medical diagnosis and legal matters, the explainability of models is crucial for ensuring fairness and trustworthiness. Research is centered around two types of interpretability (Arrieta et al. 2020): post-hoc and inherent. One benefit of post-hoc methods is that they do not hurt the performance of the original black-box models (Nielsen et al. 2022). However, the fidelity of post-hoc methods cannot be guaranteed, and the explanations can be misleading (Geirhos et al. 2024) or unreliable without context (Tomaszewska and Biecek 2024). In contrast, inherently interpretable models offer explainability through mechanism design, but their performance usually lags behind that of black-box deep learning models. The trade-off between accuracy and interpretability has been a focal point in the field of explainable artificial intelligence research (Gunning and Aha 2019; Ali et al. 2023).

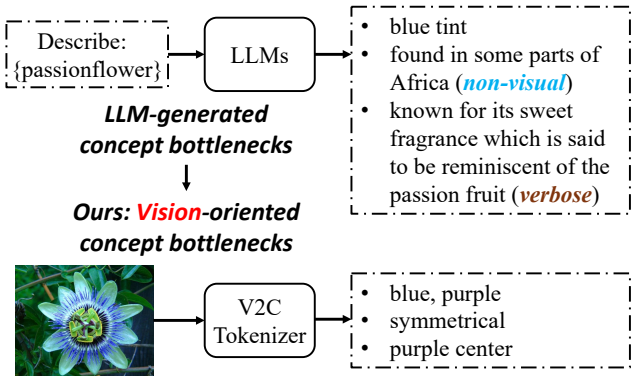

Figure 1: Problems in previous LLM-generated concept bottlenecks: non-visual and verbose concepts. Our solution: vision-oriented concept bottlenecks generated by Vision-to-Concept tokenizer directly from images.

Recently, concept bottleneck models (CBMs) have gained prominence for offering inherent interpretability with competitive performance (Koh et al. 2020). CBMs first map the image features extracted by deep learning models into a set of human-interpretable concepts (such as *red head* or *white chest* for bird classification), and then employ a linear layer to aggregate these concepts for the final prediction. The two-step design of CBMs also allows for intervention by manually altering the concept predictions. When suitable and accurate concept labels are provided, CBM can achieve comparable accuracy with better interpretability. Koh et al. and Zarlenga et al. have demonstrated the effectiveness of CBMs in fine-grained bird classification (Wah et al. 2011) and celebrity recognition (Liu et al. 2015) tasks.

Although CBMs hold promise, the annotation of concepts for visual recognition tasks requires considerable expert knowledge and labor, which impedes their widespread adoption and scalability. Recent research has addressed this challenge by using large language models (LLMs) to generate class-specific descriptions as concepts, and by harness-

*Corresponding author

ing pre-trained vision-language models (VLMs) to construct CBMs (Panousis, Ienco, and Marcos 2023; Menon and Vondrick 2023). These methods have successfully scaled CBMs to datasets of the ImageNet scale, even attaining performance on par with the original VLMs. However, as illustrated in Figure 1, the concepts generated by LLMs (such as GPT-3, Brown et al.) are obtained by directly querying the LLM with class names, which presents two issues: 1) many of the generated concepts are **non-visual**, which are hard to be captured by the vision encoder, thereby reducing accuracy and faithfulness, 2) the concepts can be **verbose** which may contain multiple attributes in one concept, and it is hard to identify the exact concepts used by the model for prediction, diminishing the interpretability of CBMs.

In this work, we propose to tackle these issues by directly generating class-specific concepts from images without the help of LLMs. To avoid verbose concepts, we use common words as our concept vocabulary and propose a concept filtering method to filter out non-visual attributes. Then a Vision-to-Concept (V2C) tokenizer is constructed using the vocabulary to quantize images into visual concepts. A contemporaneous work (Rao et al. 2024) also explores the idea of building concepts from common words, but their method requires an additional sparse autoencoder (Huben et al. 2024) trained on the large-scale CC3M dataset with labels (Ng et al. 2021) to name internal neurons as concepts. In contrast, our method can generate class-specific concepts without training, which is more efficient for resource-limited tasks like few-shot learning. We find that even without the knowledge of LLMs, our V2C tokenizer can still discover interpretable visual concepts. Our contributions can be summarized as follows.

1. We propose the V2C tokenizer to discover visual concepts directly from images, avoiding the use of LLMs.
2. We adopt common words as our concept vocabulary and develop a concept filtering method to remove non-visual and irrelevant concepts using auxiliary unlabeled images.
3. The V2C-CBM that is built on the vision-oriented concepts generated by our V2C tokenizer can achieve high classification accuracy across various datasets with visually interpretable concepts.

Related Work

VLM-based Concept Bottleneck Models

Traditional CBMs require annotated concepts for each class and training the concept predictor using these labels, which impedes the scalability of CBMs (Koh et al. 2020; Kim et al. 2023; Xu et al. 2024). Recent research has leveraged VLMs to project image features and concept texts into a shared feature space and use the cosine similarity as the concept predictor, making it more scalable for using numerous descriptions as concepts. LF-CBM is the first CBM that uses GPT-3 concepts and scales to ImageNet (Oikarinen et al. 2023), and it removes concepts that are too long or similar and then uses CLIP-Dissect (Oikarinen and Weng 2023) to filter out concepts that don't activate CLIP highly. Yang et al. propose LaBo which harnesses GPT-3 to form class-specific bottlenecks and can be used for few-shot classification, and they also employ submodular optimization (Bach 2010) for concept selection. LM4CV proposes a learning-to-search method to discover a concise set of concepts generated by LLMs (Yan et al. 2023). Res-CBM translates black-box residual vectors with unclear meanings in PCBM-h (Yüksekgönül, Wang, and Zou 2023) into potential concepts to improve performance and preserve interpretability (Shang et al. 2024). Besides CBMs, some works also adopt a similar idea of leveraging language descriptions in improving the classification accuracy of VLMs, such as CDM (Panousis, Ienco, and Marcos 2023) and DCLIP (Menon and Vondrick 2023). However, all of the above methods require a set of concepts predefined by human experts or generated by LLMs. The former may reduce the scalability of CBMs since they need human effort and expert knowledge, while the latter has issues in that the concepts generated by LLMs may be overly verbose and non-visual.

Image Quantization and Concept Discovery

Image quantization methods aim to translate images into a set of discrete tokens from a codebook, or in the context of this paper, a set of concepts. A significant amount of work in this area is based on Auto-Encoder approaches, utilizing an encoder-decoder architecture to achieve image quantization through image reconstruction task (van den Oord, Vinyals, and Kavukcuoglu 2017; Esser, Rombach, and Ommer 2021; Lee et al. 2022; Zarlenga et al. 2022; Huang et al. 2023; Zhang et al. 2023; Liu, Yan, and Abbeel 2023; Zhu, Wei, and Lu 2024; Zhu et al. 2024). The idea of using discrete codes to represent images features can also be used for extracting meaningful concepts from black-box models, which has shown to be successful in explaining LLMs (Elhage et al. 2022; Huben et al. 2024). DN-CBM (Rao et al. 2024) adopts a similar idea of quantizing image features into the most similar concepts saved in the dictionary, then building a CBM using the discovered concepts. Yet their work requires additional training of a sparse autoencoder on a large dataset to discover these concepts. In contrast, we find that with the help of a large number of unlabeled web images, we can construct a V2C tokenizer and V2C-CBM directly using the VLMs, without the need for training or LLMs.

Method

Problem Definition and Method Overview

Consider an image dataset $\mathcal{D} = \{(x, y)\}$ where x is the image and $y \in \mathcal{Y}$ is a label from N classes, and we have the class name or few-shot images x_{fs}^k for each class k, we need a set of concepts of the dataset to build a CBM.

In this work, instead of querying LLMs using class names to obtain the concept sets, we leverage an unlabeled image set $\mathcal{U} = \{x_u\}$ to construct a vision-to-concept (V2C) tokenizer $\mathcal{T}$, which can generate vision-oriented concepts directly from images. Figure 2 presents an overview of our method. First, we use class-specific features $\mathcal{F}_{base}^k$ extracted from the class name or x_{fs} of class k to select unlabeled images, which forms the quantization unlabeled image set $\mathcal{U}^q$ (Figure 2 (a)). Then, we use the most common words as

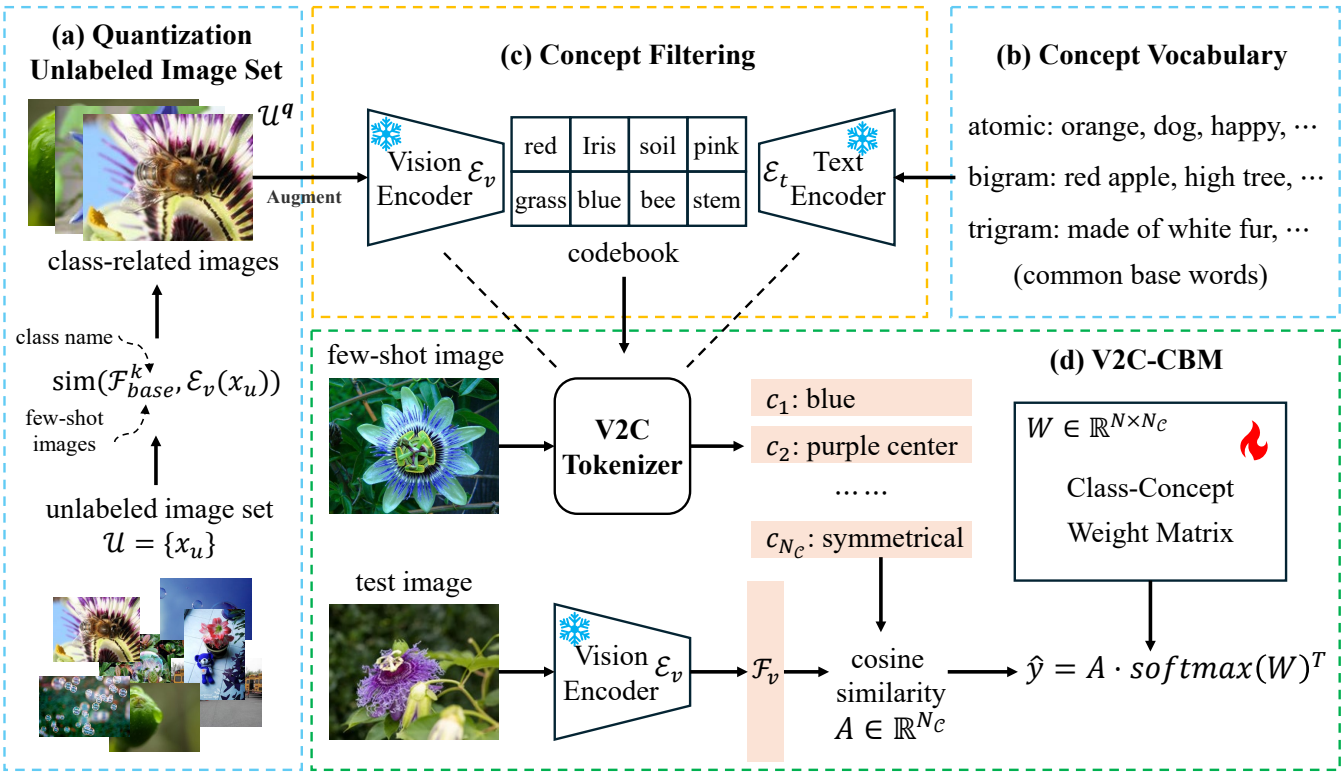

Figure 2: **Method overview**: (a) construct quantization unlabeled image set $\mathcal{U}^q$ using class-related base features $\mathcal{F}^k_{base}$ of class k, (b) adopt the most common words as the base concept vocabulary and use bigrams and trigrams to extend the vocabulary, (c) filter out non-visual and irrelevant concepts using $\mathcal{U}^q$ to form the codebook for V2C tokenizer, (d) build V2C-CBM with vision-oriented concept bottlenecks generated by V2C tokenizer from images. Our method is computation-efficient since we only need to train a class-concept weight matrix W, and don't require LLMs to discover class-specific concepts.

our base concept vocabulary (Figure 2 (b)) and propose a concept filtering method to use a VLM with vision encoder $\mathcal{E}_v$ and text encoder $\mathcal{E}_t$, and $\mathcal{U}^q$ to construct the codebook for V2C tokenizer (Figure 2 (c)). And finally, we can build V2C-CBM using the concept bottlenecks generated by the V2C tokenizer from images (Figure 2 (d)).

Quantization Unlabeled Image Set

We first build an unlabeled image set $\mathcal{U} = \{x_u\}$ to extract useful information for generating vision-oriented concepts, $\mathcal{U}$ can be obtained easily from large-scale web images. Then, we extract class-related features $\mathcal{F}^k_{base}$ for class k from the class name $\mathcal{N}_k$ using text encoder $\mathcal{E}_t$ or few-shot images x^k_{fs} using vision encoder $\mathcal{E}_v$:

$$\{\mathcal{F}^k_{base}\} = \{\mathcal{E}_t(\mathcal{P}(\mathcal{N}_k)\} \quad \text{or} \tag{1}$$

$$\{\mathcal{F}^k_{base}\} = \{\mathcal{E}_v(x^k_{fs})\} \tag{2}$$

where $\mathcal{P}$ is a set of predefined text prompts like "a photo of {class name}" to query the VLM to get more robust class features. Subsequently, we use the base features to quantize the unlabeled web images $\mathcal{U}$ into task-related $\mathcal{U}^q = \{x^q_u\}$. Specifically, we select the images with highest similarity scores from $\mathcal{U}$ for each class k to form $\mathcal{U}^q = \mathcal{U}^q_1 \cup \cdots \cup \mathcal{U}^q_k \cup \cdots \cup \mathcal{U}^q_N$, where the subscript k denotes the quantization dataset for the k-th class. The similarity is measured by the cosine similarity between the base class feature and the image feature:

$$\text{sim}(\mathcal{F}^k_{base}, \mathcal{E}_v(x_u)) = \frac{\mathcal{F}^k_{base} \cdot \mathcal{E}_v(x_u)}{\|\mathcal{F}^k_{base}\| \cdot \|\mathcal{E}_v(x_u)\|} \tag{3}$$

Concept Vocabulary

Without the knowledge of LLMs, we simply adopt common words as the base concept vocabulary. Specifically, we first collect the most common words used to describe and indicate objects in daily life as the atomic vocabulary. To further enhance the representation ability of the vocabulary set, we include some relational vocabulary (such as *part of*, *made of*, *is a*, and *has a*), and use bigrams and trigrams to expand the vocabulary set. Because directly combining all the atomic words may lead to a large vocabulary and many unreasonable concepts, we construct the bigrams by combining adjectives $\{a\}$ and nouns $\{n\}$, and construct the trigrams by combining relational words $\{r\}$, adjectives and nouns. As exampled in Figure 2 (b), the base concept vocabulary can be represented as:

$$\mathcal{C} = \{a, \cdots, n_1, \cdots\} \cup \{a_1 n_1, a_1 n_2, \cdots\} \cup \{r_1 a_1 n_1, \cdots\}$$

We also remove concepts that contain class names in the dataset to prevent information leakage.

Concept Filtering

To filter out the non-visual and irrelevant vocabulary, we use the quantization unlabeled image set $\mathcal{U}^q$ with VLM to filter the concepts. Specifically, we use the text encoder $\mathcal{E}_t$ to extract the concept features $\mathcal{F}_{concept} = \{\mathcal{E}_t(c)\}$ for $c \in \mathcal{C}$. Then, we calculate the similarity between the concept feature and the image feature to filter out the non-visual (with low similarity) and irrelevant concepts (with low frequency). In order to better align the image features of unlabeled images with fine-grained concepts, for each image x_u^q, we also generate a set of augmented images $\mathcal{A}(x_u^q)$ to extend the image set. The augmented images are generated by applying random cropping, rotation, and then resizing to the original image. The similarity between the concept feature and the augmented image is calculated as:

$$\text{sim}(\mathcal{F}^c_{concept}, x_u^q) = \frac{\mathcal{F}^c_{concept} \cdot \mathcal{E}_v(\mathcal{A}(x_u^q))}{\|\mathcal{F}^c_{concept}\| \cdot \|\mathcal{E}_v(\mathcal{A}(x_u^q))\|} \quad (4)$$

then we save M most frequent concepts for each $\mathcal{U}_k^q$ as the final concept vocabulary $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \cdots \cup \mathcal{C}_N$, the size of the final concept vocabulary will be $N_{\mathcal{C}} = M \times N$, which will be used as the codebook of V2C tokenizer. Finally, we extract and save the concept features $\mathcal{F}_{concept} = \{\mathcal{E}_t(c)\}$ for each concept c in the vocabulary as the embedding matrix.

Vision-to-Concept Tokenizer

With the concept vocabulary $\mathcal{C}$ and the saved features, we can construct the V2C tokenizer $\mathcal{T}$ to generate concepts by image quantization. Given an image x, the V2C tokenizer firstly extracts the image feature $\mathcal{F}_v$ using $\mathcal{E}_v$, then converts the image feature into top-K nearest concepts depending on the Euclidean distance between the image feature and the saved concept features:

$$\mathcal{T}(\mathcal{F}_v) = \{c_1, c_2, \cdots, c_K\} = \underset{c \in \mathcal{C}}{\arg\min} \|\mathcal{F}_v - \mathcal{E}_t(c)\|_2^2 \quad (5)$$

Given class-specific few-shot images, we can use the V2C tokenizer to generate the concepts for each class and select the most frequent concepts (depending on the size of the bottleneck) as the class-specific concept bottleneck.

V2C-CBM

Similar to other VLM-based CBMs, we use the vision encoder $\mathcal{E}_v$ and text encoder $\mathcal{E}_t$ to project the image x and the set of concepts c into a shared feature space and use cosine similarity scores A as the concept prediction, then the final prediction is made by a linear layer optimized by images and image-level labels y:

$$\hat{y} = \text{sim}(\mathcal{E}_v(x), \mathcal{E}_t(\mathcal{C})) \cdot \sigma(W)^T = A \cdot \sigma(W)^T \quad (6)$$

$$\min_W \mathcal{L}(\hat{y}, y) = \mathcal{L}(A \cdot \sigma(W)^T, y) \quad (7)$$

where $W \in \mathbb{R}^N \times \mathbb{R}^{N_C}$ is the weight matrix of the linear layer, σ is the softmax function applied along the concept axis, and $\mathcal{L}$ is the cross-entropy loss. Following Yang et al., we initialize the weight matrix W with the concept priors of the V2C tokenizer to improve the few-shot classification performance when there is very little annotated data (e.g., 1- or 2-shots learning). Specifically, if a concept c is generated by $\mathcal{T}$ using images of the k-th class, we set the corresponding elements of W as 1, otherwise 0. For cases with more labeled images, we randomly initialize the weight matrix W (more details in the ablation study section).

Experimental Setup

Datasets

We choose the following datasets for evaluation: CIFAR10, CIFAR100 (Krizhevsky and Hinton 2009), ImageNet (Russakovsky et al. 2015) as the standard benchmarks for image classification; Aircraft (Maji et al. 2013), CUB (Wah et al. 2011), Flower (Nilsback and Zisserman 2008), and Food-101 (Bossard, Guillaumin, and Gool 2014) for fine-grained image classification; DTD (Cimpoi et al. 2014) for texture classification; RESISC45 (Cheng, Han, and Lu 2017) for remote sensing scene classification; and HAM10000 (Tschandl, Rosendahl, and Kittler 2018) for skin tumor classification. We also use the same few-shot images and settings as LaBo and CLIP for a fair comparison. The classification accuracy on the test set is reported.

Implementation Details

We use CLIP ViT-L/14 to build our V2C tokenizer and V2C-CBM. For concept vocabulary, we use the English word frequency described in (Norvig 2009), and use NLTK library (Xue 2011) to determine adjectives and nouns to build the concept vocabulary. For the unlabeled image set $\mathcal{U}$, we randomly sample images from the ImageNet training set, and the default number of the unlabeled images is 200k. We use class names to extract the base features $\mathcal{F}_{base}$ for each class. $N_{\mathcal{C}}$ is set to 50 for all datasets. For each image, we select the top K (set to 5) concepts to update frequency. We then rank the word frequencies and select the top M (set to 500) words. Adam (Kingma and Ba 2015) is used for optimization, and the detailed hyperparameters are provided in the supplementary material. All experiments are conducted on an NVIDIA A100 80G PCIE graphics card using PyTorch. Since our method only requires training the class-concept weight matrix W, it is computationally efficient.

Evaluation

Baselines

We compare the classification performance of our V2C-CBM with other concept label-free methods including LaBo (Yang et al. 2023), CDM (Panousis, Ienco, and Marcos 2023), DCLIP (Menon and Vondrick 2023) and DN-CBM (Rao et al. 2024), and compare the few-shot classification performance with LaBo. The concept sets are kept the same as their original settings for a fair comparison. We report the results using the same backbone for all methods, and the linear probe (LP) performance of the black-box model is also provided for reference.

Model	Concept	CLIP ViT-L/14									
		Aircraft	CIFAR10	CIFAR100	CUB	DTD	Flower	Food	HAM	RESISC45	ImageNet
LP	-	64.0 ±0.28	98.0 ±0.02	87.5 ±0.08	84.5 ±0.10	81.5 ±0.39	99.5 ±0.01	93.2 ±0.09	82.9 ±0.36	93.9 ±0.82	83.9 ±0.09
LaBo	GPT-3	**61.3** ±0.22	97.8 ±0.11	86.0 ±0.02	81.9 ±0.03	76.9 ±0.10	**99.3** ±0.01	92.4 ±0.05	80.8 ±0.55	91.1 ±0.78	84.0 ±0.06
CDM	GPT-3	-	98.0	**86.4**	-	-	-	-	-	-	83.4
DCLIP	GPT-3	-	-	-	63.5	54.4	-	92.4	-	-	75.0
DN-CBM	SAE	-	**98.1**	86.0	-	-	-	-	-	-	83.6
Ours	-	60.7 ±0.01	98.0 ±0.03	**86.4** ±0.01	**83.0** ±0.12	**78.2** ±0.29	98.8 ±0.14	**92.8** ±0.04	**81.0** ±0.12	**92.6** ±0.26	**84.1** ±0.05

Table 1: Classification accuracy (%). LP stands for linear probing. The results of DCLIP and DN-CBM are from their respective works, and CDM is from the DN-CBM paper. The standard deviation is derived from three random experiments.

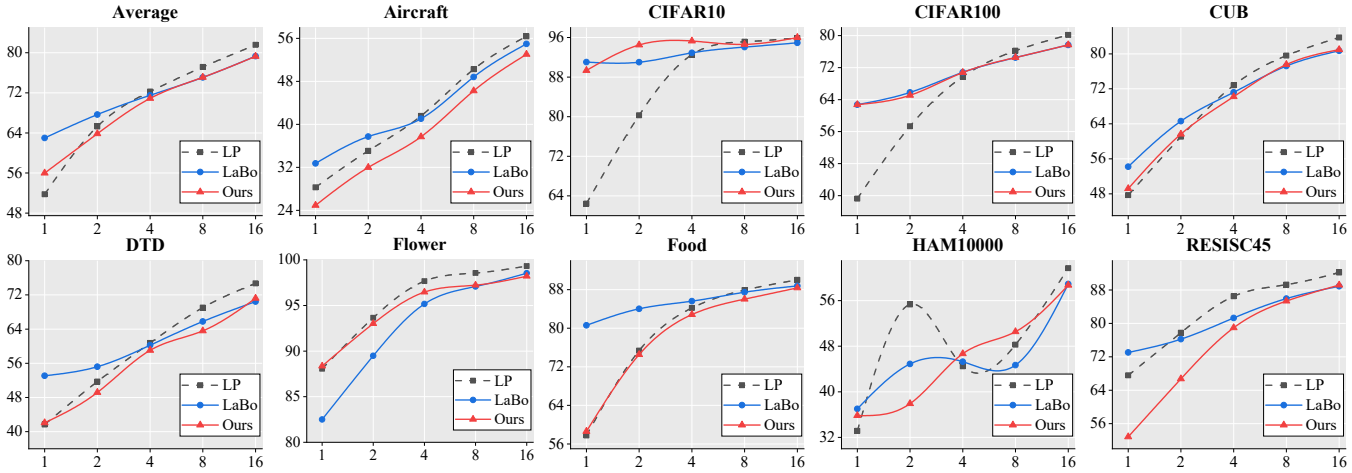

Figure 3: Few-shot classification accuracy (x-axis denotes the number of shots and y-axis denotes test accuracy).

Method	number of shots					
	1	2	4	8	16	all
LP	51.8	65.3	72.3	77.1	81.6	86.9
LaBo	**63.0**	**67.7**	**71.5**	75.1	79.3	85.2
Ours	57.8	64.0	71.1	**75.8**	**79.7**	**85.6**

Table 2: Average classification accuracy (%) on all datasets.

Classification Accuracy

In Table 1, we present the classification accuracy of our V2C-CBM on ten datasets. Our method achieves classification accuracy that is better than or comparable to the baseline methods across all datasets, without leveraging the prior knowledge of LLMs like LaBo and CDM, or additional training of a SAE for concept discovery like DN-CBM. V2C-CBM also surpasses the black-box linear probing performance of CLIP on the ImageNet dataset, showing the great scalability of our method on large datasets. Although lacking the extensive knowledge provided by LLMs or Internet encyclopedias (such as Wikipedia and WordNet), we show that making full use of the unlabeled images can also lead to competitive or even better performance, and can discover class-specific interpretable visual concepts with our method (more details in Table 3).

Few-shot Classification Performance

Figure 3 illustrates the few-shot performance of our V2C-CBM on 9 datasets. Due to the lack of extensive prior knowledge possessed by LLMs, our V2C-CBM usually performs less effectively than the GPT-3 concepts guided method LaBo when the number of labeled images is very small (1-shot and 2-shot), especially in fine-grained classification tasks like aircraft and food classification, but it can achieve comparable or slightly superior performance to the black-box linear probing method with few labeled images on almost all datasets with better interpretability.

However, the classification accuracy of V2C-CBM can increase quickly as the number of labeled images grows. As shown in Table 2, V2C-CBM achieves accuracy close to LaBo in 4-shots learning and exceeds it after 4-shots. We think this makes sense because the more labeled images, the

Class Name	V2C Tokenizer	LaBo	CDM	DCLIP
brambling 	**1.** black head **2.** brown back **3.** common bird	**1.** small, sparrow-like bird with a streaked brown back **2.** found in woods and forests across Europe and Asia **3.** found in woods and forests in Europe and Asia	**1.** barrow **2.** a long, thin, orange root **3.** large wings	**1.** a small, sparrow-like bird **2.** brown and white plumage **3.** a black head with a white stripe above the eye
hot pot 	**1.** hot bowl **2.** hot dishes **3.** red soup	**1.** circular metal object with a handle on the side **2.** round metal container with a handle on the side **3.** conical lid with a knob in the center	**1.** droopy lips and ears **2.** a game room **3.** a small, pointed tail	**1.** a pot or other container with a heating element **2.** a power cord **3.** a bowl or other vessel for holding food
ice bear 	**1.** white bear **2.** white enclosure **3.** white animal	**1.** large, white bear that lives in the Arctic **2.** perfect for keeping the bear warm in its icy habitat **3.** very important bear	**1.** white wingtips **2.** referees **3.** a note from Santa	**1.** large, white bear **2.** long neck **3.** small ears

Table 3: Top-3 concept Visualization of different methods.

more robust and accurate the concepts generated by the V2C tokenizer from images will be. For example, the *black head* concepts might not be generated by our V2C tokenizer when this part of the objects happens to be obscured in the limited few-shot images. In contrast, the advantage of LaBo lies in its ability to generate robust concept descriptions for similar categories with the help of the prior knowledge of LLMs, thus enhancing the generalization of the model.

Concept Visualization

To see whether the V2C tokenizer can discover valid visual concepts without the help of LLMs, we iterate through all the images of a particular class in the datasets and select the most frequent concepts used by our V2C tokenizer. The visualization of the top-3 concepts for some classes in the ImageNet dataset is illustrated in Table 3. Compared to other methods, our V2C tokenizer can generate concise and visually informative concepts, capturing the salient features of the target classes. For example, the *white* for ice bear and the *black head* for brambling birds. We provide more concept visualization in the supplementary material.

Ablation Study

Size of the Unlabeled Image Set

In Table 4, we investigate the impact of the size of the unlabeled image set on the final performance of V2C-CBM. As the size of $\mathcal{U}$ increases from 1k to 200k, the model's performance shows an overall increasing trend. Since a $\mathcal{U}$ of size 200k has proven to be sufficiently effective and considering the runtime efficiency, we don't continue to increase the dataset size and use 200k as our default setting.

Dataset	number of unlabeled images					
	1k	40k	80k	120k	160k	200k
CIFAR10	97.6	97.7	97.8	97.9	97.5	**98.0**
CUB	80.3	81.4	81.6	81.9	82.2	**83.0**
DTD	73.1	76.3	76.8	77.4	77.6	**78.0**
RESISC45	90.2	91.8	91.9	92.0	92.0	**92.6**

Table 4: Classification accuracy (%) using unlabeled image sets with different number of images.

Bigram and Trigram Concepts

We also investigate the impact of different combinations of conceptual vocabulary on model performance. Specifically, we explore the scenarios of using only atomic concepts (A), using both atomic and bigram concepts (AB), and employing atomic, bigram, and trigram concepts simultaneously (ABT). The experimental results are presented in Table 5. We find that the model achieves better performance in the ABT scenario. Looking back at the discovered concepts shown in Table 3, we think the combination is beneficial for describing similar objects with more accurate and distinct attributes (e.g., *white fur* v.s. *fur*). This may also explain why LaBo can exhibit stronger few-shot capabilities in fine-grained classification tasks besides the knowledge provided by LLMs — while concepts generated by LLMs are more complex, assigning diverse concepts to distinguish similar categories becomes easier. However, when the number of labeled images is sufficiently large, the advantage of using complex concepts is no longer pronounced, particularly in the context of our vision-oriented concept bottleneck.

Dataset	Type	number of shots					
		1	2	4	8	16	all
Food	A	52.3	73.2	81.7	85.7	88.0	92.0
	AB	57.7	73.9	82.0	85.8	88.0	92.1
	ABT	**58.6**	**74.5**	**82.5**	**86.3**	**88.9**	**92.8**
RESISC45	A	60.5	68.2	78.4	84.7	88.4	91.6
	AB	61.9	69.8	78.1	84.3	88.4	91.5
	ABT	**62.0**	**69.8**	**80.7**	**86.6**	**89.7**	**92.6**

Table 5: Classification accuracy (%) using different concept combination method: A denotes atomic words, B denotes bigrams and T denotes trigrams.

Base Feature Type	number of shots				
	1	2	4	8	16
class name	**88.4**	**93.0**	**96.5**	**97.2**	**98.2**
few-shot images	77.2 (11.2↓)	86.6 (6.4↓)	93.8 (2.7↓)	96.4 (0.8↓)	97.7 (0.5↓)

Table 6: Classification accuracy (%) on the Flower dataset using different class-related base features.

Vocabulary Set	number of shots					
	1	2	4	8	16	all
base	**58.6**	**74.5**	**82.5**	**86.3**	**88.9**	**92.8**
without food	29.7 (28.9↓)	58.3 (16.2↓)	74.5 (8.0↓)	80.3 (6.0↓)	83.2 (5.7↓)	86.3 (6.5↓)
only food	50.2 (8.4↓)	71.9 (2.6↓)	81.2 (1.4↓)	85.3 (1.0↓)	87.7 (1.2↓)	92.4 (0.4↓)

Table 7: Classification accuracy (%) on the Food dataset using different vocabulary sets.

Method	number of shots					
	1	2	4	8	16	all
LP	51.8	65.3	72.3	77.1	81.6	87.0
LaBo	63.0	67.7	71.5	75.1	79.3	85.3
Ours$_p$	**57.8**	**64.0**	70.3	74.7	78.7	85.2
Ours$_r$	50.4	61.6	**71.1**	**75.8**	**79.7**	**85.7**

Table 8: Average classification accuracy (%) on all datasets using different initialization methods. Ours$_p$ means initializing with concept priors and Ours$_r$ means random.

Different Types of Class-related Features

We also examine the discrepancy in extracting class-related features $\mathcal{F}_{base}^k$ by employing class names versus using few-shot images. For class names, we utilize a collection of 85 prompt templates, leveraging the text encoder to extract features and computing the mean of these features to serve as the base feature $\mathcal{F}_{base}$. The specifics of the text prompts are detailed in the supplementary material; for few-shot images, we compute the mean of the extracted image features as the base feature. The performance on the Flower-102 dataset is illustrated in Table 6. We observe that the use of class names and text prompts demonstrates better robustness, particularly when the number of labeled images is limited. The gap between the text features derived from class names and the image features diminishes as the number of labeled images increases. This is consistent with our previous understanding of the few-shot performance, which is the more images, the more robust the concepts generated by the V2C tokenizer.

Different Vocabulary Set

We also test the impact of different sources of concept vocabulary on model performance. In Table 7, we try, respectively: 1) using the top 10k most common English words (base); 2) removing words that can be used to describe food in the base vocabulary, then using the rest of the words as another vocabulary, which produces a vocabulary unrelated to the classification task (without food); and 3) using the filtered out words as the vocabulary (only food). It can be observed that when using vocabulary unrelated to the task (without food), the model's performance significantly drops, with an accuracy reduction of up to 28.9% in 1-shot tasks and 16.2% in 2-shot tasks. The base vocabulary achieves similar but higher accuracy than the only food vocabulary, which indicates that our method can effectively discover relevant concepts to the target task from the large vocabulary.

Initialization with Priors

Following Yang et al. (2023), we also test whether using concept priors in the V2C tokenizer (the language priors in LaBo) can improve the classification performance. As shown in Table 8, initializing with concept priors is more effective for very limited labeled images (1- and 2-shots), yet random initialization is more powerful as the number of labeled images grows. The conclusion is similar to LaBo's, which states that prior is more important for low shot settings since there is less signal to guide concept importance. So we use concept priors to initialize W for 1- and 2-shot learning while using random initialization for the others.

Conclusion

In this work, we construct a vision-oriented concept bottleneck without the reliance on LLMs by developing a V2C tokenizer that maps images to a discrete set of concepts. The creation of the V2C tokenizer necessitates only a set of unlabeled images, which can be readily acquired from the Internet. We show that the V2C tokenizer can discover interpretable visual concepts and lead to V2C-CBM, which has surpassed LLM-guided CBMs across various datasets and even outperformed black-box linear probing methods on the ImageNet dataset, showcasing the efficacy of our method. Follow-up work may be devoted to developing methods for evaluating the trustworthiness of VLM-based CBMs with open-vocabulary concepts.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 62394314, 82371112, 623B2001, 62394311, and in part by the High-grade, Precision and Advanced University Discipline Construction Project of Beijing (BMU2024GJJXK004).

References

Ali, S.; Abuhmed, T.; El-Sappagh, S. H. A.; Muhammad, K.; Alonso-Moral, J. M.; Confalonieri, R.; Guidotti, R.; Ser, J. D.; Rodríguez, N. D.; and Herrera, F. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion*, 99: 101805.

Arrieta, A. B.; Rodríguez, N. D.; Ser, J. D.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58: 82–115.

Bach, F. 2010. Convex Analysis and Optimization with Submodular Functions: a Tutorial. arXiv:1010.4207.

Bossard, L.; Guillaumin, M.; and Gool, L. V. 2014. Food-101 - Mining Discriminative Components with Random Forests. In *ECCV (6)*, volume 8694 of *Lecture Notes in Computer Science*, 446–461. Springer.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

Cheng, G.; Han, J.; and Lu, X. 2017. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE*, 105(10): 1865–1883.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing Textures in the Wild. In *CVPR*, 3606–3613. IEEE Computer Society.

Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; Grosse, R.; McCandlish, S.; Kaplan, J.; Amodei, D.; Wattenberg, M.; and Olah, C. 2022. Toy Models of Superposition. arXiv:2209.10652.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*, 12873–12883. Computer Vision Foundation / IEEE.

Geirhos, R.; Zimmermann, R. S.; Bilodeau, B.; Brendel, W.; and Kim, B. 2024. Don't trust your eyes: on the (un)reliability of feature visualizations. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 15294–15330. PMLR.

Gunning, D.; and Aha, D. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI magazine*, 40(2): 44–58.

Huang, M.; Mao, Z.; Chen, Z.; and Zhang, Y. 2023. Towards Accurate Image Coding: Improved Autoregressive Image Generation with Dynamic Vector Quantization. In *CVPR*, 22596–22605. IEEE.

Huben, R.; Cunningham, H.; Riggs, L.; Ewart, A.; and Sharkey, L. 2024. Sparse Autoencoders Find Highly Interpretable Features in Language Models. In *ICLR*. OpenReview.net.

Kim, E.; Jung, D.; Park, S.; Kim, S.; and Yoon, S. 2023. Probabilistic Concept Bottleneck Models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, 16521–16540. PMLR.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *International conference on machine learning*, 5338–5348. PMLR.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4).

Lee, D.; Kim, C.; Kim, S.; Cho, M.; and Han, W. 2022. Autoregressive Image Generation using Residual Quantization. In *CVPR*, 11513–11522. IEEE.

Liu, H.; Yan, W.; and Abbeel, P. 2023. Language Quantized AutoEncoders: Towards Unsupervised Text-Image Alignment. In *NeurIPS*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *ICCV*, 3730–3738. IEEE Computer Society.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-Grained Visual Classification of Aircraft. arXiv:1306.5151.

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

Ng, E. G.; Pang, B.; Sharma, P.; and Soricut, R. 2021. Understanding Guided Image Captioning Performance across Domains. In *CoNLL*, 183–193. Association for Computational Linguistics.

Nielsen, I. E.; Dera, D.; Rasool, G.; Ramachandran, R. P.; and Bouaynaya, N. C. 2022. Robust Explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Process. Mag.*, 39(4): 73–84.

Nilsback, M.; and Zisserman, A. 2008. Automated Flower Classification over a Large Number of Classes. In *ICVGIP*, 722–729. IEEE Computer Society.

Norvig, P. 2009. Natural language corpus data. *Beautiful data*, 219–242.

Oikarinen, T. P.; Das, S.; Nguyen, L. M.; and Weng, T. 2023. Label-free Concept Bottleneck Models. In *ICLR*. OpenReview.net.

Oikarinen, T. P.; and Weng, T. 2023. CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks. In *ICLR*. OpenReview.net.

Panousis, K. P.; Ienco, D.; and Marcos, D. 2023. Sparse Linear Concept Discovery Models. In *ICCV (Workshops)*, 2759–2763. IEEE.

Rao, S.; Mahajan, S.; Böhle, M.; and Schiele, B. 2024. Discover-then-Name: Task-Agnostic Concept Bottlenecks via Automated Concept Discovery. In *ECCV (77)*, volume 15135 of *Lecture Notes in Computer Science*, 444–461. Springer.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.*, 115(3): 211–252.

Shang, C.; Zhou, S.; Zhang, H.; Ni, X.; Yang, Y.; and Wang, Y. 2024. Incremental Residual Concept Bottleneck Models. In *CVPR*, 11030–11040. IEEE.

Tomaszewska, P.; and Biecek, P. 2024. Position: Do Not Explain Vision Models Without Context. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 48390–48403. PMLR.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1).

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural Discrete Representation Learning. In *NIPS*, 6306–6315.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Xu, X.; Qin, Y.; Mi, L.; Wang, H.; and Li, X. 2024. Energy-Based Concept Bottleneck Models: Unifying Prediction, Concept Intervention, and Probabilistic Interpretations. In *ICLR*. OpenReview.net.

Xue, N. 2011. Steven Bird, Evan Klein and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc 2009. ISBN: 978-0-596-51649-9. *Nat. Lang. Eng.*, 17(3): 419–424.

Yan, A.; Wang, Y.; Zhong, Y.; Dong, C.; He, Z.; Lu, Y.; Wang, W. Y.; Shang, J.; and McAuley, J. J. 2023. Learning Concise and Descriptive Attributes for Visual Recognition. In *ICCV*, 3067–3077. IEEE.

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. In *CVPR*, 19187–19197. IEEE.

Yüksekgönül, M.; Wang, M.; and Zou, J. 2023. Post-hoc Concept Bottleneck Models. In *ICLR*. OpenReview.net.

Zarlenga, M. E.; Barbiero, P.; Ciravegna, G.; Marra, G.; Giannini, F.; Diligenti, M.; Shams, Z.; Precioso, F.; Melacci, S.; Weller, A.; Lió, P.; and Jamnik, M. 2022. Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off. In *NeurIPS*.

Zhang, J.; Zhan, F.; Theobalt, C.; and Lu, S. 2023. Regularized Vector Quantization for Tokenized Image Synthesis. In *CVPR*, 18467–18476. IEEE.

Zhu, L.; Wei, F.; and Lu, Y. 2024. Beyond Text: Frozen Large Language Models in Visual Signal Comprehension. In *CVPR*, 27037–27047. IEEE.

Zhu, L.; Wei, F.; Lu, Y.; and Chen, D. 2024. Scaling the Codebook Size of VQGAN to 100,000 with a Utilization Rate of 99 arXiv:2406.11837.