Cross-Modality Image Interpretation via Concept Decomposition Vector of Visual-Language Models

Zhengqing Fang, Zhouhang Yuan, Ziyu Li, Jingyuan Chen, Kun Kuang, Yu-Feng Yao, and Fei Wu, *Senior Member, IEEE*

***Abstract*— Interpretable image classification is crucial for making decisions in high-stakes scenarios. Recent advancements have demonstrated that interpretable models can achieve performance comparable to black-box models by integrating Visual Language Models (VLMs) with Concept Bottleneck Models (CBMs). These models explain their predictions by calculating the weighted sum of similarities between the image representation and predefined text embeddings. However, selecting textual descriptors is subjective, and relying solely on textual information may not capture the complexities of visual data, impacting both interpretability and performance. To address these limitations, this work explores the cross-modality interpretation of class-related concepts in image classification. Specifically, we propose decomposed concept bottleneck model (DCBM), which utilizes a set of decomposed visual concepts that are extracted directly from images instead of predefined text concepts. The decomposition of concepts is achieved through vector projection onto concept decomposition vectors (CDVs), which can be interpreted across both textual and visual modalities. We introduce a quintuple notion of concepts and a concept-sample distribution theorem, which enables the localization of decomposed concepts in images using the Segment Anything Model (SAM) with automatically generated prompts. Experimental results demonstrate that DCBM achieves competitive performance compared to non-interpretable models, with a 3.42% improvement in classification accuracy and a 66.27% improvement in image-text groundability compared to other VLM-based CBMs. Furthermore, we evaluate the benefits of employing automatically generated prompts in SAM for interpreting visual concepts, in contrast to prompts created by human operators.**

***Index Terms*— Explainable artificial intelligence, concept bottleneck models, segment anything, visual-language models.**

I. Introduction

INTERPRETABLE models are characterized by their ability to present a decision-making process that is intelligible to humans. They are particularly useful in high-stakes decision-making situations, owing to their transparency and capacity for explanation [1], [2], [3]. One specific type of interpretable model is Concept Bottleneck Models (CBMs) [4], which employ a two-step prediction process as shown in Figure 1. First, they predict an intermediate set of concepts (c) that are specified by humans, along with their corresponding concept scores (s). Then, these scores are used to predict the final classification output (y) via a weighted sum. However, a significant limitation of CBMs is the necessity for extensive annotated concepts during the training phase. To reduce the labeling effort, recent works combine CBMs with visual language models (VLMs) [5], as the VLMs shows remarkable zero-shot classification performance solely relying on natural language text. They calculate concept scores as the similarities between image representations and embeddings of encoded pre-defined texts, facilitating easy expansion in concept numbers and enhancing classification performance [6].

However, current VLM based CBMs might present two issues. Firstly, the selection of an appropriate set of textual concepts is subjective, and manual construction may not cover all relevant visual features for a certain class. To address this, one possible solution is to utilize Large Language Models (LLMs) to generate textual concepts [7], [8], [9]. Nevertheless, it is important to note that this approach may introduce **unreliable concepts** that are difficult to validate, potentially compromising the reliability of the model's interpretation. An example of failure is shown in Figure 2, where the concept set of "*Blue Grosebeak*" does not include important clues such as the "*blue wing*", while unreliable concepts such as non-factual, non-visual and task irrelevant descriptions are present. Secondly, a phenomenon called **modality gap** [10] has consistently been observed across various VLMs, which potentially undermine the model's downstream zero-shot classification

Manuscript received 27 December 2023; revised 8 April 2024; accepted 11 May 2024. Date of publication 20 May 2024; date of current version 7 April 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFE0204200, in part by the National Natural Science Foundation of China under Grant U20A20387 and Grant 62376243, and in part by the National Youth Foundation of China under Grant 62307032. This article was recommended by Associate Editor H. Zhao. *(Zhengqing Fang and Zhouhang Yuan contributed equally to this work.) (Corresponding authors: Jingyuan Chen; Yu-Feng Yao; Fei Wu.)*

Zhengqing Fang and Zhouhang Yuan are with the Department of Ophthalmology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310016, China, and also with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China.

Ziyu Li is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China.

Jingyuan Chen is with the College of Education, Zhejiang University, Hangzhou 310005, China (e-mail: jingyuanchen@zju.edu.cn).

Kun Kuang is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China, and also with the Key Laboratory for Corneal Diseases Research of Zhejiang Province, Hangzhou 310016, China.

Yu-Feng Yao is with the Department of Ophthalmology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310016, China, and also with the Key Laboratory for Corneal Diseases Research of Zhejiang Province, Hangzhou 310016, China (e-mail: yaoyf@zju.edu.cn).

Fei Wu is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China, and also with the Key Laboratory for Corneal Diseases Research of Zhejiang Province, Hangzhou 310016, China (e-mail: wufei@cs.zju.edu.cn).

Digital Object Identifier 10.1109/TCSVT.2024.3403167

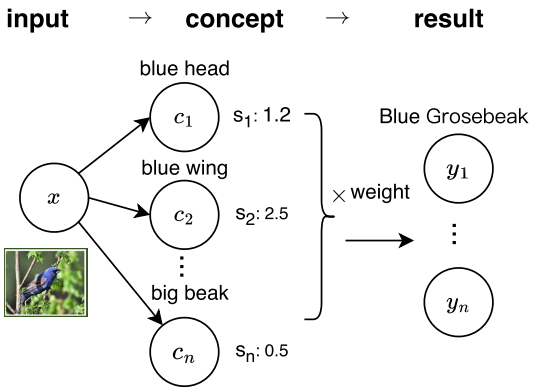


Fig. 1. The basic workflow of concept bottleneck models for inherent interpretable classification.

performance. Specifically, as shown in Figure 2, a noticeable gap can be observed between the visual embedding space (highlighted in green) and the textual embedding space (highlighted in orange). This gap has a negative impact on the accuracy of the predicted concept scores in CBMs and results in a potential inconsistency between input image and output texts that used to interpret the classification result, which may hurt the users' trust of their interpretation.

To address the aforementioned limitations, we propose a novel approach called Decomposed Concept Bottleneck Model (DCBM). Unlike existing methods that rely on predefined textual concepts, DCBM introduces decomposed visual concepts instead. This helps avoid the incorporation of unreliable textual concepts and also addresses the modality gap issue. As shown in Figure 2, the input image of a blue bird is classified by comparing it with decomposed visual concepts such as *"blue and black head"* (s_3) and *"blue wings"* (s_4), rather than relying on textual concepts generated by LLMs that are associated with the category label *"Blue Grosbeak"* (s_1 and s_2). Specifically, the decomposed visual concepts are represented by Concept Decomposition Vectors (CDVs) which has the same size as image representations. The classification process of DCBM is similar to existing CBM methods, while the key difference is that concept scores are the similarity between pretrained image representation and CDVs instead of text embeddings.

Building DCBM poses a key challenge in training and interpreting the CDVs. In the training process of CDVs, we constrain their distribution to be similar to the distribution of training image representations in an adversarial manner, and optimize CDVs to minimize classification error. This ensures that the visual semantics of CDVs are maintained, as they closely resemble the visual representations of real training samples. For the interpretation of CDVs, we innovatively perform their interpretations in a cross-modality manner. In the textual modality, we rely on vector similarity to identify the most similar text to explain the semantics of CDVs. This is made possible by the adversarial training that constrains the distribution of CDVs in the origin latent space of VLM. In the visual modality, our objective is to visually represent the semantics present in the training samples. However, the visual representation derived from the final layer of the image encoder, such as the Vision Transformer of CLIP, is a compressed vector that includes positional information but cannot be readily interpreted as a spatial arrangement of the image. To overcome this limitation and capture location information, we utilize the representation obtained from an intermediate layer of the encoder. In this layer, the features are represented as a sequence of tokens that correspond to the original image patches. We employ a neural network, namely **reverse modality converter** (θ^{-1}), to predict the representation of the semantics of CDVs in the token feature space. Consequently, we can easily generate a similarity heatmap, where regions with higher similarity are likely to contain closer semantics of the given CDVs. Moreover, the heatmap can be converted into a bounding box and points as automatic prompt (auto-prompt) to obtain more precise image segments using Segment Anything Model (SAM) [11]. This allows for the visualization of concepts in images and their expression in natural language simultaneously.

Given that concepts are abstract mental objects, individuals often recall them through concrete instances [12]. To enhance the comprehensiveness of a concept, it is beneficial to provide multiple samples that illustrate it. To describe the relationship between concepts and samples, we introduce the concept-sample distribution (CSD), which can be viewed as a categorical distribution. The CSD is obtained by measuring the similarity between the concept vector and sample embeddings, allowing us to identify the most similar samples that can effectively interpret the semantic of a concept. Therefore, each concept can then be formalized using five key elements (*i.e.*, quintuple notion): a concept decomposition vector (CDV), a class name, a scalar weight, a set of image fragments, and a set of text. In this study, we discover that the contrastive visual-language pretraining loss of CLIP aims to minimize the discrepancy between the CSDs in the textual and visual modalities. We generalize this property to the latent space of intermediate layers of vision encoder, as the training objective of the **reverse modality converter** is to transfer the CDV semantic into the intermediate latent space [13].

Experiments on various datasets demonstrate that DCBM performs competitively in terms of classification when compared to black-box classifiers. Furthermore, it surpasses existing state-of-the-art VLM-based CBMs by an average accuracy improvement of 3.42%. The quality of cross-modality interpretations is showcased through case studies at the concept, sample and class levels, which aid in comprehending the model's decision-making process. To assess the textual interpretation quality of DCBM in comparison to other VLM-based CBMs, a human evaluation is conducted [14]. Participants are asked to sort the interpreted texts given by different CBMs according to their assessment in terms of four metrics including *Groundability* (consistency between the interpreted image contents and text descriptions), *Meaningfulness* (semantic coherence of interpreted texts), *Factuality* (consistency between the interpreted texts and predicted labels), and *Fidelity* (degree of support from the concept scores for the model's predictions). The result of human study indicates that the cross-modality interpretation by DCBM outperforms current VLM-based CBMs by 66.27% in terms of Groundability,

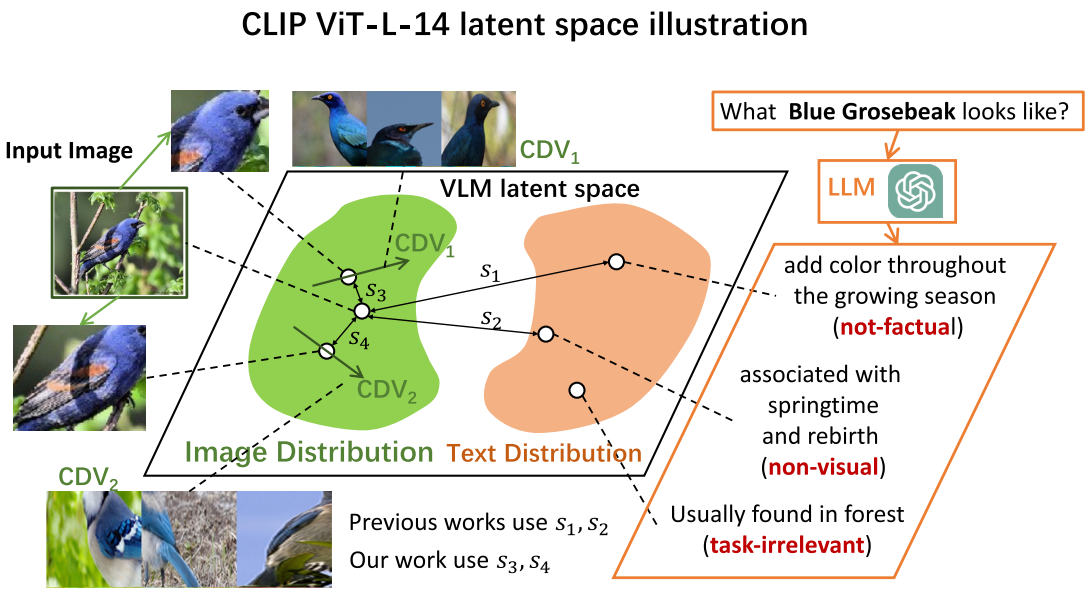


Fig. 2. Schematic of concepts from different modalities in VLM latent space. Visual concepts are expected to be used for building concept bottleneck models rather than text concepts given by LLM in this work.

40.57% in Factuality, 34.94% in Meaningfulness and 34.41% in Fidelity. Furthermore, the quality of visual interpretation is evaluated by comparing the segmentation results with ground-truth masks. This evaluation demonstrates the enhanced quality achieved through the utilization of SAM and provides insights into the instances where automatic prompts outperform human-constructed prompts within SAM. The contributions of this paper can be summarized as follows:

- Propose Concept Decomposition Vectors (CDVs) to decompose class-related visual concepts. By using CDVs, the proposed DCBM achieves competitive performance with black-box models and surpasses existing VLM-based CBMs.
- Propose the concept-sample distribution and quintuple notions for a novel form of cross-modality interpretation to explain decomposed concepts. This interpretation provides more informative explanations of classification for given datasets compared to single modality explanations and helps reveal classification challenges.
- Explore the utilization of Segment Anything Model in visual concept interpretation, where the prompt is generated automatically according to the interpreted visual concept. Experiments demonstrate the superior performance of auto-prompt over human-constructed prompt.

II. Related Works

A. Explanation Methods in Deep Learning

The explanation of deep learning models can be categorized into post-hoc methods and inherent interpretable (*i.e.*, ad-hoc) methods. The former assumes the reasoning process is a black-box and requires the creation of a model to reveal it, including utilizing saliency maps [15], generating heatmaps [16], testing with Concept Activation Vector (CAVs) [17], providing counterfactual examples [18] or analyzing the network's response to image perturbations [19]. Post-hoc methods do not interfere with the architecture, but may produce biased and unreliable explanations [1]. Consequently, there is increasing focus on designing ad-hoc models that make the decision process directly visible. For example, ProtoPNet [20] employs a prototype hidden layer that represents activation patterns. Concept Bottleneck Models (CBMs [4]) force the intermediate layers to be semantic via multi-label supervised learning. As the inherent interpretable models are more reliable, they are applied in various tasks like 3D points could [21], medical image [22], [23] and marketing [24].

Recent development of ad-hoc methods includes deep nearest centroid (DNC) [25], which utilizes sub-centroids to describe class distributions and explains classifications based on the proximity of test data to these sub-centroids in the feature space. ProtoConcepts [26] enhances prototype-based networks by using multiple visualizations to represent prototypical concepts. Deformable ProtoPNet [27] uses spatially flexible prototypes to capture pose variations and context, improving accuracy and explanation richness. Despite the innovations, they predominantly focus on the single visual modality and may not fully capture the semantic meaningfulness that is significant to human understanding compared to our DCBM. Note that the Concept Decomposed Vectors (CDVs) used in DCBM is similar to the sub-centroids of DNC but with a key distinction: CDVs are trained to encapsulate decomposed visual concepts, whereas sub-centroids represent the visual representation of an entire image. As a result, we use additional techniques to explain CDV in both textual and visual modalities.

B. Concept Bottleneck Models

Concept Bottleneck Models (CBM) [4], [28] is a class of inherent interpretable models. A concept bottleneck model learns a mapping from samples to labels in two steps: (i) a

concept encoder function which maps samples from the input space to an intermediate concept space, and (ii) a label predictor function which maps samples from the concept space to a downstream task space. A CBM requires a dataset composed of tuples in *samples* × *concepts* × *labels*, where each sample consists of an input image with label and multiple ground truth concept annotations. The original CBM [4] represent concepts as boolean and fuzzy univariates that aligned with a single ground truth concept or represent a probability of that concept being active. However, such representation suffers from a trade-off between interpretability and classification accuracy [29]. Therefore, Concept Embedding [28] are proposed to alleviate this trade-off, who learn two embeddings per concept, one for when it is active, and another when it is inactive.

Building above CBMs requires exhausted concept annotation effort, which makes CBM less applicable in real-world application [6]. With the development of VLM, researchers of Post-hoc CBM [6], Label-free CBM [7], and LaBo [8] integrated images and texts and used the similarity between them as concept scores for classification, named as VLM-based CBMs in this paper. However, these VLM-based models directly use text embedding rather than concept embedding to represent a concept, suffering the **modality gap** and **unreliable concept** (see Section I). Our DCBM also integrate VLMs to enhance CBMs but distinguishes itself from previous VLM-based CBMs by avoiding pre-defined concept texts. Instead, we learn concept representations from the training images and interpret them after training. By doing so, DCBM avoids the information loss caused by **modality gap** and achieves improvements in both classification and interpretation.

C. Visual-Language Models and Prompt Engineering

Visual Language Models (VLMs) are pre-trained using vast image-text pairs readily available on the internet, enabling direct application to visual recognition tasks without fine-tuning [30], as demonstrated by CLIP [5], which leverages the contrastive loss to align paired images and texts in embedding space. CLIP's notable zero-shot transferability allows it to perform diverse tasks without supervision, such as image retrieval [31], zero-shot classification [5], and image generation [32]. Subsequent models like BLIP2 [33] and GLIP [34] have expanded on CLIP's cross-modal capabilities. The zero-shot ability of CLIP relys on the visual encoder's image representations and the language encoder's text embeddings, where the input texts are sentences containing target class name. However, the similarities is reported to be influenced by the modality gap [10]. Specifically, the latent spaces of visual and textual embeddings are limited in narrow cones, leading to a gap between the latent spaces of two modalities. This gap are affected by the hyperparameter in the contrastive loss during training process. To date, how to close the gap remains to be unsolved [10] and its impact on finer-grained correspondence between vision and language has space to explore [35]. Thus, direct utilizing image-text similarities to explain model decision might be unsatisfying.

Another concern is that the image-text similarities are highly sensitive to the texts, e.g. the choice of prompts. A prompt is a template text; for instance, *"**a photo of** a dog"* yields a very different similarity score from *"a dog"* [5]. This sensitivity to texts might be a crucial factor that existing VLM-based CBMs fails to output consistent zero-shot classification [36]. To enhance classification performance, a line of researches have tried to construct the most suitable prompt for specific tasks (prompt engineering) or automatically learn prompts embedding [37], [38] (prompt tuning). This can be achieved by adding learnable tokens [39], introducing learnable layers between models [40], incorporating low-rank additional parameters [41], or adding learnable tokens around samples [42]. While our DCBM also includes learnable parameters, it differs from prompt tuning because our objective is to identify and interpret the visual information inherently present in the samples, whereas the prompt tuning method is solely focused on performance enhancement for a specific task.

D. Segment Anything Model

Segment Anything Model (SAM) [11] is a new paradigm for semantic segmentation that is trained with more than one billion ground-truth segmentation masks in 11 million natural images. It can produce high-quality segmentation results based on different types of prompts, including positive/negative (foreground/background) points, a rough box or mask, or freeform text. However, an empirical study conducted across different scenarios [43] has revealed that SAM exhibits a limited understanding of specialized data. It often requires strong prior knowledge to craft effective prompts. In response to this challenge, a body of research has emerged, focusing on tailoring SAM for specific application scenarios [44]. This includes its application in fields such as remote sensing [45] and medical image analysis [46].

In this work, we harness SAM to segment concept regions within images based on automatically generated prompts derived from concept activation heat-maps. By doing so, we can obtain a more refined segmentation of the region in question, which in turn allows us to validate the effectiveness of the concept decomposition vector. This approach not only leverages the power of SAM but also integrates the strengths of our interpretive mechanisms, leading to a more nuanced and precise understanding of the visual concepts present.

III. Preliminaries

A. Contrastive Image-Text Matching Pretraining of VLMs

Visual Language Models (VLMs) are a series of models that can understand and generate both images and text, *e.g.*, CLIP [5]. Image text matching (ITM) is the common training objective that maps image representation into a language concept embedding space. We formalize the ITM training objective in the next paragraph.

Let $x \in \mathcal{X}$ denote an image x in an image set $\mathcal{X}$, and $t \in \mathcal{T}$ represent a text t in a text set $\mathcal{T}$. $\{(x_i, t_i)|i = 1, \ldots, N\}$ denotes N image-text pairs, the match relationship can be represented with an identical matrix $\mathbf{Y}$, where

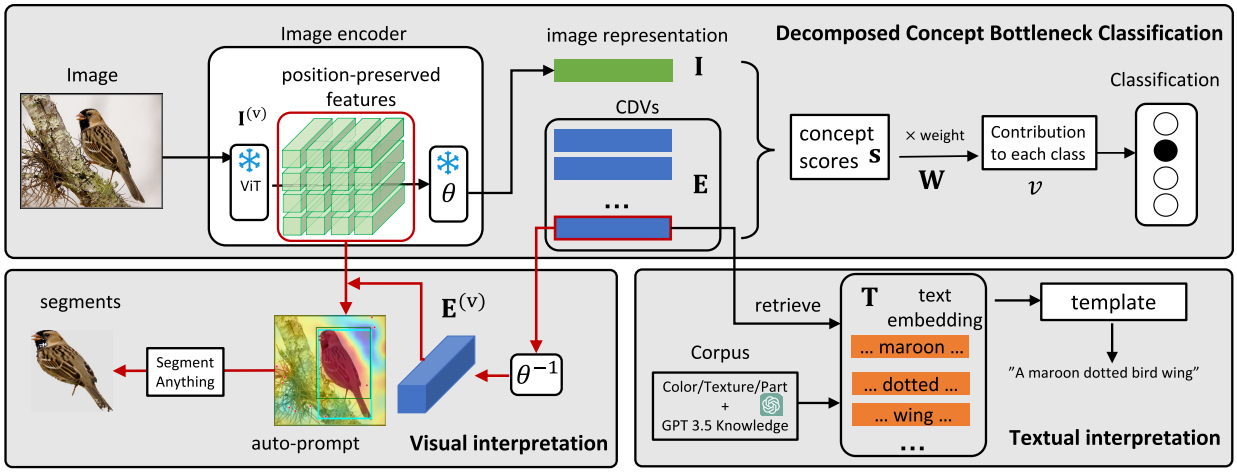


Fig. 3. The workflow of decomposed concept bottleneck model (DCBM) with well-trained CDVs. The classification process is shown in the upper part, where an image embedding is projected to CDVs and the classification result is the weighted sum of vector similarities (concept scores). The lower part exhibit the cross-modality interpretation by taking one CDV for example. For visual interpretation, the CDV is fed into reverse modality converter θ^{-1} to predict its representation in the intermediate latent space. This allows the CDV to be visualized through a similarity heatmap with position-preserved image tokens. Additionally, the heatmap can be fed into Segment Anything Model as an automatic prompt for finer visualization. For textual interpretation, the CDV computes the concept-sample distribution (CSD) with multiple candidate text sets $\mathcal{T}_c$ to identify the closest texts. These closest texts can then be alternatively composed into a sentence.

$\mathbf{Y}_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}$. In general, a VLM consists of an image encoder $I(\cdot)$, which maps the input image x into a d-dimensional embedding space $\mathbb{R}^d$, and a text encoder $T(\cdot)$ which maps the input text t into $\mathbb{R}^d$. We can get an image embedding matrix $\mathbf{I} = [I(x_1), \ldots, I(x_N)]$ and a text embedding matrix $\mathbf{T} = [T(t_1), \ldots, T(t_N)]$, where $\mathbf{I}, \mathbf{T} \in \mathbb{R}^{N \times d}$. The model is trained to maximize the similarity between the embeddings of matching image and text pairs as:

$$\min_{I,T} \left[H(\sigma(\tfrac{\mathbf{I}\cdot\mathbf{T}^\top}{\tau}), \mathbf{Y}) + H(\sigma(\tfrac{\mathbf{T}\cdot\mathbf{I}^\top}{\tau}), \mathbf{Y}) \right], \tag{1}$$

where σ is the softmax operation applied in each row, $H(\cdot,\cdot)$ is the cross-entropy function $H(p,q) = -\sum_i p(i) \log q(i)$, and τ is a learnable temperature coefficient. After training, VLMs become zero-shot classifiers by computing dot product between embeddings of input image x_0 and candidate text $t_i \in \mathcal{T}_0$, where $\mathcal{T}_0 = \{t_i | i = 1, \ldots, K\}$ is the set of K text combining the i-th category names with prompt texts. The zero-shot classification probability of class k given x_0 is $p(k|x_0) = \sigma(\frac{I(x_0)\cdot\mathbf{T}_0^\top}{\tau})_k$, where $\mathbf{T}_0 = [T(t_1), \ldots, T(t_K)]$ is the embedding matrix of $\mathcal{T}_0$.

B. VLMs in Concept Bottleneck Models

Given an input image x_0, a CBM predicts a concept score s_c for each human-readable concept c via a shared neural network $f(\cdot)$, and get bottleneck concept scores $\mathbf{s} = [s_0, s_1, \ldots, s_C]$. C is the number of predefined concepts. The final model decision is obtained by multiplication with sparse weight matrix $\mathbf{W} \in \mathbb{R}^{C \times K}$ for a K-classification problem as:

$$\mathbf{p}(k|x_0) = \sigma(\mathbf{W}^\top \cdot \mathbf{s})_k, \text{ where } \mathbf{s} = f(x_0). \tag{2}$$

This prediction score of class k, denoted as v_k, can be interpreted as $v_k = \sum \mathbf{s}_i \times \mathbf{W}_{ik}$, where i is the index of concepts. Recently, LaBo [8] and Lable-Free CBM [7] combine VLMs with CBM to reduce the exhausting labeling effort of $f(\cdot)$. They directly use image encoder I to replace f and get a set of concept text $\mathcal{T}_c$ from LLMs. In this case, the bottleneck concept scores $\mathbf{s}$ are obtained as:

$$\mathbf{s} = \frac{I(x_0) \cdot \mathbf{T}_c^\top}{\tau}, \tag{3}$$

where $\mathbf{T}_c = [T(t_1), \ldots, T(t_C)]$ is the embedding matrix of $\mathcal{T}_c$. By this means, the label effort issue is addressed to some extent, and the model performance seems to be improved with respect to the increased number of concepts.

In this paper, we claim that the choice of $\mathcal{T}_c$ is subjective and the bottleneck concept scores $\mathbf{s}$ is not accurate because of modality gap. For better interpretability and performance, we intend to directly obtain a visual concept embedding matrix $\mathbf{E}$ to compute bottleneck scores $\mathbf{s} = \frac{I(x_0)\cdot\mathbf{E}^\top}{\tau}$ and explain the visual concept embedding in both vision and language. Each embedding $\mathbf{e}_i$ is a d-dimensional vector named as concept decomposition vector (CDV), where d is the dimension of image embedding $I(x_0)$. The CBM with CDV is named as decomposed concept bottleneck models (DCBM).

In this paper, we argue that the selection of $\mathcal{T}_c$ is subjective, and the bottleneck concept scores $\mathbf{s}$ are inaccurate due to the modality gap. To enhance interpretability and performance, we propose directly obtaining a visual concept embedding matrix $\mathbf{E}$. This matrix allows us to calculate bottleneck scores $\mathbf{s}$ as $\mathbf{s} = \frac{I(x_0)\cdot\mathbf{E}^\top}{\tau}$ to avoid the modality gap, and subsequently explained in both visual and textual modalities. Each embedding $\mathbf{e}_i$ in $\mathbf{E}$ is a d-dimensional vector, termed a Concept Decomposition Vector (CDV), with d corresponding to the dimension of the image embedding $I(x_0)$. We refer to the CBM utilizing CDVs as the Decomposed Concept Bottleneck Model (DCBM).

C. Overall Procedure of DCBM

DCBM contains three main modules as shown in Figure 3. The classification module involves training CDV with frozen image encoders and a discriminator in an adversarial manner, while simultaneously minimizing classification loss using a weighted classifier. Leveraging CDVs allows our method to achieve a dual interpretation of both visual and text modalities. For visual interpretation, we employ the SAM model as an external dependency to determine the regions of concepts within the image. This is accomplished by calculating the correlation between the CDVs corresponding to the sample and the latent space representation of the sample, prompting SAM to perform concept positioning and segmentation. In text interpretation, we utilize LLM and VLM as external dependencies. We build a text set through LLM and then match concepts and CDVs through VLM. Finally, we complete the semantic matching of CDV. Our method not only establishes a conceptual basis for classification but also provides explanations by indicating the activation positions of concepts in the image and their corresponding semantics.

IV. Concept Decomposition Vector

A concept decomposition vector (CDV), denoted as $\mathbf{e}$, is a vector in the VLM latent space $\mathbb{R}^d$, which captures some key visual concepts that distinguish a class from others. We can combine multiple concepts into a concept matrix $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n]$. With well-trained $\mathbf{E}$, we can perform Decomposed Concept Bottleneck Model (DCBM) by taking place $\mathbf{T}_c$ in Eq. 3 for inherent interpretable classification. The workflow is shown in the upper part of Figure 3, where both encoders I and T are frozen. Section IV-B outlines how to learn CDVs from a given training dataset, and Section IV-A introduces the quintuple notion of concept and concept-sample distribution. To make each $\mathbf{e}$ as well as its representing concept human-understandable, cross-modality interpretation is performed as described in Section IV-C.

A. Quintuple Notion and Concept-Sample Distribution

Definition 1 (Quintuple notion of concept): Each $\mathbf{e}$ is assigned to one category a with a scalar weight $w \in \mathbb{R}$, representing the concept is decomposed from the category. As concepts are mental objects, to let someone realize the concept, a set of image patches $\mathbb{I}$ and a set of text phrases $\mathbb{T}$ needs to be exhibited at the same time. Therefore, the concept represented by CDV $\mathbf{e}$ is denoted as a quintuple so the concept matrix $\mathbf{E}$ contains a set of decomposed visual concepts:

$$\mathcal{E} = \{(\mathbf{e}_i, a_i, w_i, \mathbb{T}_i, \mathbb{I}_i)\}_{i=1}^{N}. \tag{4}$$

The CDV $\mathbf{e}$, assignment a, and weight w are determined in Section IV-B. The image set $\mathbb{I}$ and text set $\mathbb{T}$ of Eq. 4 are determined given CDV $\mathbf{e}$ in Section IV-C. To determine the image set $\mathbb{I}$ and text set $\mathbb{T}$ given CDV $\mathbf{e}$, we need another definition of the concept and sample relationship.

Definition 2 (Concept-sample distribution): Given a sample set $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ and a concept embedding $\mathbf{e} \in \mathbb{R}^d$, the concept-sample distribution (CSD) is defined as a categorical distribution over the sample set $\mathcal{Z}$ with following probability density function:

$$\delta(k; \mathbf{e}, \mathcal{Z}) = \frac{\exp(\mathbf{e}\cdot\mathbf{z}_k)}{\sum_{\mathbf{z}\in\mathcal{Z}} \exp(\mathbf{e}\cdot\mathbf{z})}, \tag{5}$$

where $\mathcal{Z}$ can either be a text set or an image set. For convenience, we denoted CSD as $\delta(\mathbf{e}, \mathcal{Z})$.

Proposition 1: The pretraining task of VLMs [5], contrastive image-text matching, is to minimize two concept-sample distributions with a shared concept $\mathbf{e}_i$ between different modalities sample set $\mathcal{I}$ and $\mathcal{T}$ given image-text pair (x_i, t_i). Formally, Eq. 1 is equivalent to the following objective:

$$\min_{\mathbf{I},\mathbf{T}} \sum_i^N [\mathrm{KL}(\mathbf{Y}_i \| \delta(\mathbf{e}_i, \mathcal{T})) + \mathrm{KL}(\mathbf{Y}_i \| \delta(\mathbf{e}_i, \mathcal{X}))]. \tag{6}$$

The proof is trivial by setting the concept embedding $\mathbf{e}$ as the text embedding $\mathbf{t}_l$ and the sample set $\mathcal{Z}$ as the image embedding $\mathbf{x}_l$ in Eq. 5. This motivates us to use the concept-sample distribution as a learning objective to train a concept embedding in arbitrary latent space.

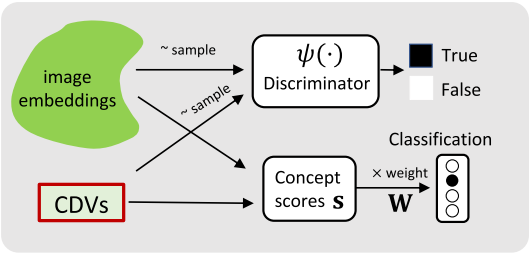


Fig. 4. Adversarial training of CDVs.

B. Learning Process of Concept Decomposed Vector

1) Initialization: Given training image dataset with labels $\mathcal{D} = (x_i, y_i)$. We calculate mean $\mu_{\mathcal{X}}$ and variance $\sigma_{\mathcal{X}}$ of image features. Let C be the number of CDVs, $\mathcal{C}$ be a categorical distribution with equal probability $\bar{p}$, and $\mathcal{U}$ be a uniform distribution. The first three terms of quintuple are initialized as $\mathcal{E} = \{(e_i, a_i, w_i) | e_i \sim \mathcal{N}(\mu_{\mathcal{X}}, \sigma_{\mathcal{X}}), a_i \sim \mathcal{C}(\bar{p}), w_i \sim \mathcal{U}(0, 1)\}$. Then we get concept matrix $\mathbf{E}$, and sparse weight matrix $\mathbf{W}$ with w_i as elements on the one-hot embedding of a_i. As we hope that the CDV itself represents a visual concept, we constrain the distribution of CDVs to be consistent with the visual concepts that appeared in the training set. To achieve this, we apply adversarial training to learn CDVs as shown in Figure 4. There are two steps in each iteration:

2) Step one (train discriminator): A random initialized 3-layer neural network with non-linear activation $\psi(\cdot)$ acts as a discriminator to tell CDV $\mathbf{e}$ from real image feature $I(x_i)$. In each iteration, we first sample a batch of CDVs $\{\mathbf{e}_i\}$ and a batch of image features $\{\mathbf{z}_i | \mathbf{z}_i = I(x_i)\}$. Then we calculate the loss of discriminator $\mathcal{L}_\psi$ as follows:

$$\mathcal{L}_D = -\frac{1}{m} \sum_{i=1}^{m} \big[\log \psi(\mathbf{e}_i + \epsilon) + \log(1 - \psi(\mathbf{z}_i + \epsilon))\big]. \tag{7}$$

where $\epsilon \sim N(0, 1)$ is sampled from standard Gaussian noise in each iterations for robust training.

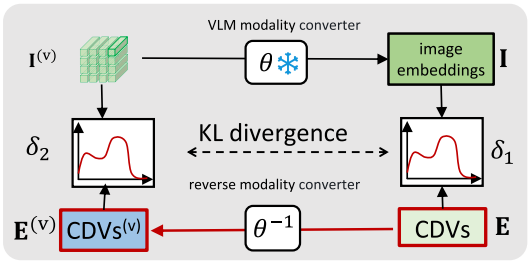


Fig. 5. Learning concepts representation in vision space.

3) Step Two (Train CDVs): We use the discriminator to train CDVs to be indistinguishable from real image features. At the same time, we perform interpretable classification by taking $\mathbf{E}$ into Eq. 3. Jointly train $\mathbf{E}$ and $\mathbf{W}$ with negative log-likelihood loss. The final loss of step two $\mathcal{L}_{CDV}$ is defined as follows:

$$\mathcal{L}_{CDV} = \underbrace{-\frac{1}{m}\sum_{i=1}^{m}\log\psi(\mathbf{e}_i+\epsilon)}_{\text{discriminator loss}} + \underbrace{\frac{1}{|\mathcal{X}|}\sum y_i \log(\sigma(\frac{I(x_i)\cdot\mathbf{E}^{\top}}{\sqrt{\eta}}\cdot w^*))}_{\text{classification loss}} + \underbrace{\mathcal{R}(\mathbf{E})}_{\text{regularizer}}. \tag{8}$$

The regularization loss $\mathcal{R}(\mathbf{E}) = \sum_i \sum_j \frac{\mathbf{e_i}*\mathbf{e_j}}{\|\mathbf{e_i}\|\|\mathbf{e_j}\|}$ constrains the CDVs in $\mathbf{E}$ to be as orthogonal as possible with each other.

C. Cross-Modality Interpretation of CDV

1) Language Comprehension via Text Composition: Given a CDV $\mathbf{e}$, the set of text $\mathbb{T}$ is sampled from the CSD $\delta(\mathbf{e}, \mathcal{T}_c)$. $\mathcal{T}_c$ is constructed with two strategies with the aid of GPT3.5 [47]: (1) *category-related sentences*: for general class names that LLMs are familiar with, we use a prompt combined with the category name to get description of visual contents following previous work [7]. Then $\mathbb{T} = \{t_i | t_i \sim \delta(\mathbf{e}, \mathcal{T}_c)\}$. (2) *category-independent words*: for fine-grained class names that LLMs are easily confused, use human defined corpus to get category-independent words $\mathcal{T}_p$ for the primitive visual concepts, e.g. colors, shapes. Then $\mathbb{T} = \bigcup_p \{t_i | t_i \sim \delta(\mathbf{e}, \mathcal{T}_p)\}$, and a text composition is optional to organize the words into a sentence by predefined templates.

2) Vision Comprehension via Auto-Prompt Segmentation: In VLMs, positional information of images is lost after being embedded by $I(\cdot)$, which makes it difficult to locate the decomposed concept in the image. To address this, we view I as two parts: a ViT that outputs position-preserved embedding $\mathbf{I}^{(v)}$ and a modality converter [48] θ that map $\mathbf{I}^{(v)}$ to $\mathbf{I}$. Each sample in $\mathbf{I}^{(v)}$ includes $1 + L \times L$ tokens, where $\mathbf{I}_0^{(v)}$ denotes the class token. Then we train a neural network θ^{-1} to act as a reverse modality converter to predict the concept representation of $\mathbf{e}$ in ViT output space. Motivated by Eq. 5, the concept representation in another embedding space can be learned by minimizing the KL divergence between the CSDs of the same concept across different modalities. Figure 5 illustrate the learning objective to predict pre-layer representation. Let $\delta_1 = \delta(\mathbf{e}, \mathbf{I})$ and $\delta_2 = \delta\left(\theta^{-1}(\mathbf{e}), \mathbf{I}_0^{(v)}\right)$, the training objective is shown as follows:

$$\theta_0^{-1} = arg\min_{\theta^{-1}} [\mathrm{KL}(\delta_1 \| \delta_2) + \mathrm{KL}(\delta_2 \| \delta_1)]. \tag{9}$$

After training, the pre-layer representation of CDVs $\mathbf{E}^{(v)}$ are predicted via $\mathbf{E}^{(v)} = \{\theta_0^{-1}(\mathbf{e}_1), \theta_0^{-1}(\mathbf{e}_2), \ldots, \theta_0^{-1}(\mathbf{e}_C)\}$. Then similarity heatmaps can be calculated between the image embeddings $I_{L\times L}^{(v)}$ and each pre-layer representation of CDV. The heatmap is scaled to the same size as the input image during visualization, and then fused with the input image to obtain the final visualized regions of interest.

Automatic prompt: In order to get finer visualized regions of interest, we use SAM [11] to further process the heatmap. Specifically, we draw a bounding box as positive prompt around the region of top 50% of the similarity in the heatmap, and take the lowest similarity position as the negative prompt. Then we use SAM to segment the input image with the automatic prompts, and finally take the mask with the highest confidence score as the fine-grained segmentation of the interpreted concept.

V. Example Interpretations of DCBM

A. Three Levels of Interpretation

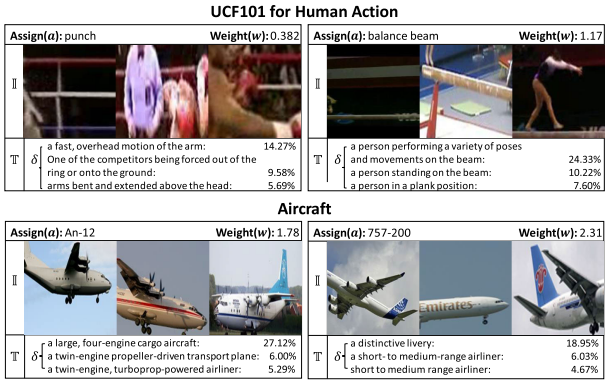
DCBM can perform 3-level interpretation, including concept, sample, and class level.[1]

1) Concept-Level Interpretation: Concept-level interpretation is the basis of DCBM. By checking whether the concept is interpreted consistently in different modalities, the DCBM users can judge whether the visual language model has learned the category-related semantics well. For each CDV, we visualize three image patches crop by the bounding box and the texts that is closest in candidate corpus to represent its semantic. Figure 6a illustrates the cross-modality understanding of the selected CDVs from UCF101 and Aircraft datasets with the category-related text descriptors. In the first example of UCF101, a concept related to the 'punch" has a weight of 0.382. The most closest texts is "*a fast, overhead motion of the arm*" with 14.27% probability. If there exists inconsistency between different modalities, it indicates that the vision and language are not well aligned in the chosen VLM. The misalignment is difficult to be detected by the sole-modal interpretation of previous VLM-CBMs.

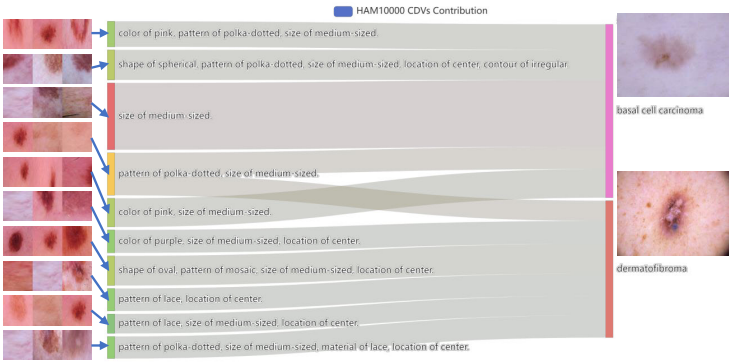
2) Class-Level Interpretation: CBMs can be explained as a linear combination of interpretable features, where the weights can be regarded as their importance for classification. We created a Sankey diagram to visualize the final layer weights. Figure 6b shows that DCBM cannot distinguish *"basal cell carcinoma"* from *"dermatofibroma"* well because they share common visual concepts that interpreted as *"polka-dotted and medium size"*. The width of the lines connecting a concept to an output class represents their weight, with only weights greater than 0.05 included. As fine-grained datasets demand specific domain knowledge for accurate classification, class-level explanations can help understand these classification challenges and provide guidance for model improvement.

3) Sample-Level Interpretation: In Figure 6c, the horizontal bar charts depict the values obtained by multiplying concept scores with their weights, where each concept is explained cross-modally. The concept can be found in the input image

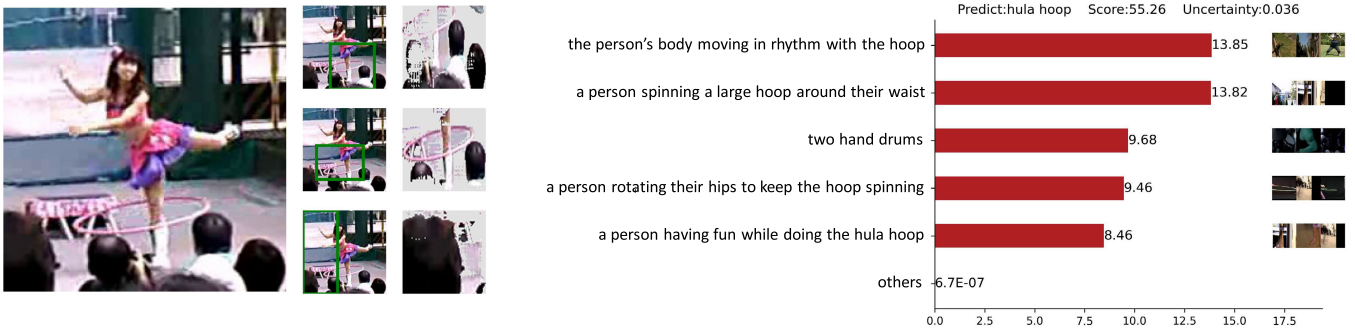
[1]All interpretations use ViT-L/14 as the image encoder of CLIP.



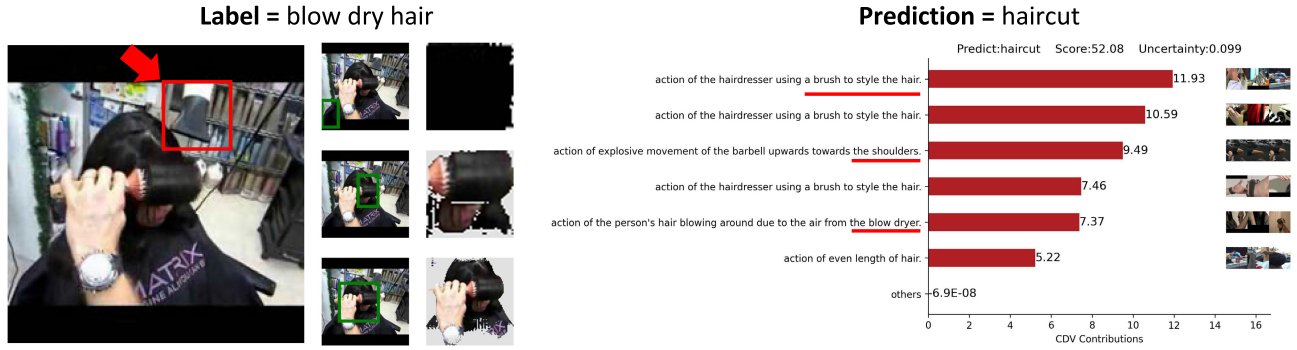
(a) Concept-level interpretation of CDVs. The visualization of each CDV contains all five elements in its quintuple notion together with concept-sample distributions.



(b) Class-level interpretation shows how DCBM differentiate the classes, where the width of the lines connecting a concept to an output class represents their weight, with only weights greater than 0.05 visualized.



(c) Sample-level interpretation from UCF101. DCBM says that the image is classified as "hula loop" because of the concept scores from "the person body" and "a large hoop" that bounded by green box (segmented by SAM [11]).



(d) Failure case interpretation from UCF101. The image labeled as *"blow dry hair"* is mistakenly classified as *"haircut"*. Through interpretation, DCBM indicates that the concepts such as *"using a brush to style the hair"* and *" the shoulders"* are more influential in the misclassification than the concept of *"the blow dryer"*. Visually, the concept of *"the blow dryer"* (manually enclosed in a red rectangle) is not identified as it is partially obscured and does not appear in its entirety.

Fig. 6. Showcases of cross-modal interpretation from (a) concept-level, (b) class-level, (c) sample-level and (d) failure prediction.

with bounding boxes and segmentation masks from SAM [11]. In this case, the concept *"the person's body moving in rhythm with the hoop"* has a contribution of 13.86 and *"a person spinning a large hoop around their waist"* has 13.82, which is consistent with the image content. The remaining concepts work as the same. The predicted score to label *"hula loop"* is 55.26 and exactly the same as the sum of the concept scores. This indicates that DCBM interprets its decision with fidelity. As other VLM-based CBMs can also provide sample-level interpretation (with only scores and texts), we compare the quality of the interpretation in the Section VI-B.

B. Interpretations of Failure Cases

Examples of failures and their interpretations are essential for researchers to gain a deeper understanding of the decision-making process and enhance model performance. In Figure 6, a failure case is presented where an input image labeled as *"blow dry hair"* is classified as *"haircut"*. The DCBM effectively explains that this failure is a result of the incomplete key concept *"the blow dryer"*. Through interpretation, we identify three main types of failure predictions: (i) Out-of-distribution samples, where concepts in the samples are challenging to identify and only a few concepts are

TABLE I
CLASSIFICATION ACCURACY OF FOUR VLM-BASED CBMS ON DIFFERENT *test* SETS

Backbone	Dataset Type	Natural	Semantic		Fine-grained		Specialized				Avg.
	Dataset Name	CIFAR100	DTD	UCF101	CUB	Aircraft	EuroSAT	HAM10000	DR	Kera	
ViT-B-16	Linear probe*	75.62%	77.89%	86.59	77.50%	52.31%	96.04%	80.64%	53.22%	68.01%	74.20%
	Sparse LP	59.84%	74.86%	80.21%	63.87%	42.67%	92.15%	76.32%	50.55%	62.59%	67.01%
	Label Free CBM	58.79%	69.05%	77.29%	59.35%	35.57%	92.21%	72.16%	**53.80%**	49.10%	63.04%
	LaBo	73.93%	75.18%	**85.67%**	76.48%	**51.48%**	93.87%	80.30%	47.51%	50.13%	70.51%
	DCBM(Ours)	**75.37%**	**77.39%**	85.60%	**77.36%**	50.25%	**95.44%**	**80.74%**	52.21%	**67.77%**	**73.57%**
ViT-L-14	Linear probe*	80.87%	81.17%	90.12%	84.15%	62.42%	97.22%	80.66%	55.10%	67.14%	77.65%
	Sparse LP	74.29%	80.54%	87.23%	81.24%	58.03%	95.70%	78.55%	**53.63%**	66.66%	75.07%
	Label Free CBM	46.54%	66.84%	74.46%	56.42%	29.49%	74.65%	70.05%	53.42%	44.43%	57.37%
	LaBo	79.62%	77.30%	**90.11%**	81.90%	**61.06%**	95.82%	**81.39%**	48.48%	44.44%	73.35%
	DCBM(Ours)	**80.86%**	**81.12%**	89.57%	**83.95%**	60.67%	**96.89%**	81.05%	52.73%	**67.23%**	**77.12%**

* Linear probe is uninterpretable, which can be viewed as a ceil performance of interpretable models in previous research.

activated significantly. (ii) Suspiciously correlated concepts, such as the visual concepts *"grass"* and *"soccer"*, leading to misclassifications of images of *"baseball"* as *"soccer"*. (iii) Boundary cases, where classes share too many common visual concepts, like *'haircut"* and *"blow dry hair"* sharing visual concepts like *'haircut" and "shoulders"*, making it challenging to assign high concept scores to discriminative concepts. It is believed that cross-model explanations can assist model users in attributing failure cases accurately and implementing suitable strategies to enhance model performance.

VI. EXPERIMENT

The experiments are conducted for the following goals: (1) to compare the classification performance of using CDV rather than the text concepts from LLM across multiple image domains. (2) to compare the interpretation with existing CBM methods in terms of text and concept scores with both automatic evaluation and human evaluation. (3) to evaluate the effectiveness of SAM in visual concept segmentation. (4) to evaluate the impact of each component via an ablation study of DCBM. [2]

A. Classification Performance Analysis

1) Datasets: (1) **Natural** images to evaluate general classification performance, using the well-known image dataset CIFAR-100 [49]; (2) **Semantic** images with clear concepts in their class names, including DTD [50], a texture dataset containing 47 human-recognizable textures, and UCF101 [51], a human actions dataset with 101 human actions; (3) **Fine-grained** images that require some additional knowledge. CUB-200-2011 [52] without cropping, a bird dataset containing 200 different classes of birds, and FGVC-Aircraft [53], an aircraft dataset containing 100 different classes; (4) **Specialized** images from real-world applications with special camera. EuroSAT [54], a satellite remote sensing image dataset containing 10 kinds of land use types, HAM10000 [55], a medical image dataset containing 7 kinds of skin diseases, Diabetic Retinopathy [56], a dataset containing 5 types of diabetic retinopathy, and Keratitis [57], a slit-lamp image dataset containing 4 kinds of infectious keratitis diseases.

[2]In all tables, bold indicates that the value is the best for the column.

2) Baseline: Four methods are chosen, including black-box linear probe [58], sparse linear probe [59] for sparse layers have been demonstrated to be more interpretable [7], Label-free CBM [7], and LaBo [8]. None of these methods alter the image encoder parameters. The performance of linear probe serves as the benchmark for interpretable methods. For the choice of VLM, all methods employ the same pre-trained CLIP model with ViT-B/16 and ViT-L/14 as image backbones. The same *train/dev/test* split with [8] is setted and select the best validation performance on *dev*, reporting the average classification accuracy of five runs with random seeds 41-45. The classification results on testsets are shown in Table I.

3) Compared to Linear Probe: Table I indicates that DCBM shows comparable performance to the linear probe on most datasets, including natural images, semantic images, and specialized images. This suggests CDV has sufficiently utilized the embedded feature. Moreover, our approach offers the added benefit of interpretability compared to linear-probe by the CBM-like classifier.

4) Compared to VLM-Based CBM: DCBM outperforms two VLM-based CBM methods, LaBo and Label-free CBM, on most datasets, particularly for natural images. On semantic images, DCBM improves significantly over both methods on DTD and on UCF101 compared to Label-free CBM, with only a marginal difference of 0.31% in average classification accuracy compared to LaBo. For specialized images, DCBM shows a significant improvement over both methods on EuroSAT and Kera datasets, but performs marginally lower than LaBo by 0.1% on average in HAM10k dataset. However, on the DR dataset, DCBM and LaBo perform worse than Label-free CBM by 1.14%, indicating a potential benefit of utilizing an external CNN feature extractor. On fine-grained images, DCBM significantly outperforms LLM-CBMs on CUB, while trailing LaBo on aircraft, potentially due to non-visual text from LLMs that could compromise interpretability.

B. Interpretability Comparison With Other CBM Models

1) Quality Comparison: Figure 7 compares the interpretation of Label-free CBM, LaBo and DCBM. In the explanation of DCBM, the corresponding text and image explanation are given for each concept, and the user can check their consistency to justify if the DCBM really understands the

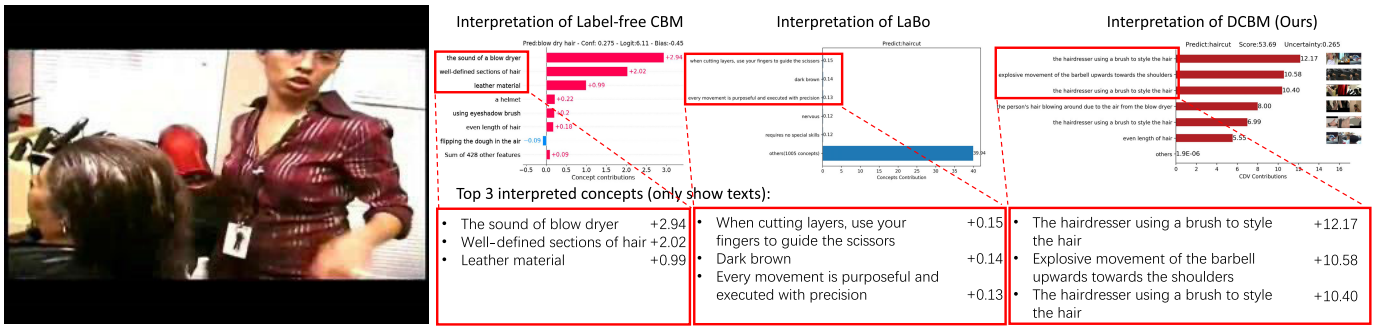


Fig. 7. Interpreting the same image (left), different VLM-CBMs gave the correct prediction but output different explanations. Label-free CBM outputs a non-visual description *"the sound of blow dryer"*, and the third-ranked text, *"leather material"*, is not shown in the image. The users may be confused by the relationship between *"haircut"* and *"leather material"* because the model's understanding of *"leather material"* may be inconsistent with humans, which cannot be justified using text only. In LaBo's explanation, 1005 concepts are used, and the poor sparsity makes it not available in practical applications.

TABLE II
DEFINITION OF EVALUATION METRICS FOR INTERPRETABILITY EVALUATION

Metric	Type	Definition	Description	Range
Accuracy	automatic	$(TP+TN)/(TP+TN+FP+FN)$	Average classification accuracy of DTD and UCF101 datasets	$[0,1]$
Sparsity	automatic	$\text{avg}\left(\sum_{i=0}^{n} s_i \times \mathbf{W}_{ik}/v_k\right)$	Average proportion of top5 concept scores to the final classification logits	$[0,1]$
Groundability	manual	$\frac{\sum 3\times N_1+2\times N_2+3\times N_3+0\times N_{n/a}}{3\times N}$	Which interpretation texts are more consistent with the image content?	$[0,1]$
Factuality	manual		Which interpretation texts are consistent with the label?	$[0,1]$
Meaningful	manual		Which interpretation texts are more semantic ?	$[0,1]$
Fidelity	manual		Which interpretation scores are more supportive to the predictions?	$[0,1]$

text. Then the interpreted score can be validated by checking whether the concept appears in the given image.

2) Quantitative Comparison: To verify which method's interpretation results are more in line with human perception, an evaluation is conducted via an online questionnaire, where 8 cases are randomly sampled from UCF101 and DTD. The testers are asked to rank the interpretation from DCBM, LaBo and Label-free CBM in terms of four metrics for each case. Participants (n=27) are invited for the diverse geographic distribution to make our research as representative as possible.

3) Metrics: (1) *Groundability* [8] assesses the alignment between interpreted texts and image content. (2) *Factuality* [8] measures the consistency between the interpreted text and the actual label. (3) *Meaningfulness* [60] evaluates whether the concept is comprehensible to humans. (4) *Fidelity* [61] determines if the interpretation scores align with the predictions. Participants rank the methods based on each metric, and we tally the frequency of top rankings for each method, with N_1 representing the number of times a method is ranked first, and $N_{n/a}$ indicating instances where a method was not included in the ranking. The metric calculations are detailed in Table II, which also includes the average *Accuracy* on DTD and UCF101 datasets to assess the balance between performance and interpretability. We also calculate *Sparsity* to automatically evaluate the ratio of the top 5 concept scores to all concepts, as users may prefer not to review an extensive list of concepts for each prediction.

4) Result: As shown in Figure 8, our method (CDV) exhibits significant improvements across six metrics when compared to LaBo and Label-free CBM. These results indicate that our interpretations are more readily accepted by

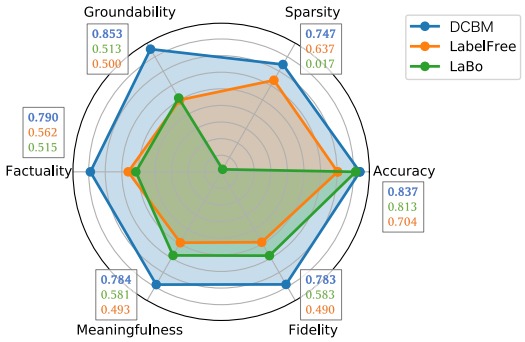
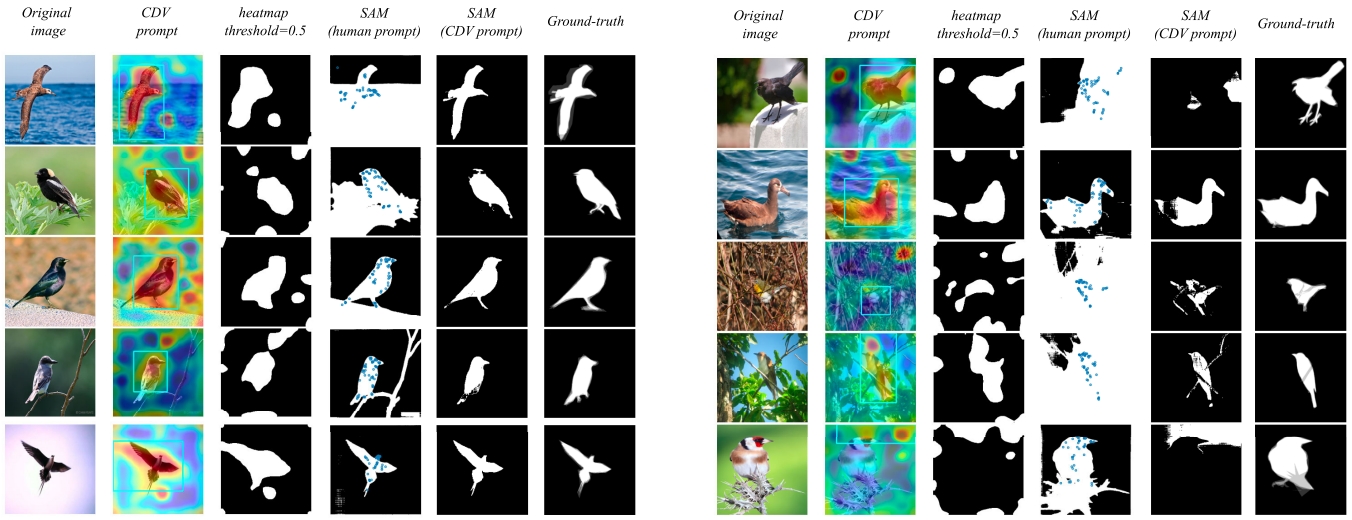


Fig. 8. Radar chart of interpretability evaluation with 6 metrics.

humans, with a notable 0.203 increase in *Meaningfulness* over the next best method. Additionally, our explanations align more closely with human understanding, as evidenced by a *Factuality* score that is 0.275 higher than that of Label-free CBM.

In terms of the *Sparsity* metric, LaBo registers significantly lower values than the other methods, despite its *Accuracy* being on par with our approach. This is attributed to LaBo's tendency to assign high weights to an excessive number of concepts for a single sample, a practice that is not advisable for real-world applications. Our CDV-DCBM effectively balances this issue, as CDV captures visual information that holds equivalent or greater value than textual information. This is further supported by the strong evidence provided by our scores of 0.853 in *Groundability* and 0.783 in *Fidelity*, indicating that our interpretations are more readily recognizable and understandable than those of other methods.



(a) Randomly sampled examples from correct predictions.

(b) Failure cases where DCBM output wrong classification results.

Fig. 9. Segmentation results of different prompts in visual concept segmentation for SAM [11] in birds classification.

C. Evaluation of SAM for Visual Concept Segmentation

In DCBM, the visual concepts could be visualized with similarity heat maps. Subsequently, we process the heat maps to obtain bounding boxes, which are then fed into the Segmentation Anything Model (SAM) to achieve more precise segmentation results. In this section, we present experiments aimed at evaluating the quality enhancement of the SAM technique and comparing the quality of automatically generated prompts (auto-prompt) with human annotated prompts.

1) Setup: The experiments are conducted with CUB200 [52]. The dataset provides human annotations of key visual concepts. For each image, five annotators record key visual concept locations by clicking on the head, feet, wings, and other relevant parts, if they are visible in the image. These human annotated points can serve as prompt for SAM. To evaluate the segmentation results, we employ the Intersection over Union (IoU) as the segmentation metric, which measures the consistency between segmentation results and ground truth. Higher IoU values indicate better segmentation alignment with actual annotations. The mean IoU value of all test sets are denoted as *mIoU*.

$$mIoU = \sum_i \frac{TP_i}{TP_i + FP_i + TN_i}, \tag{10}$$

where TP_i is the number of true positive pixels of the i-th test sample, FP_i is the number of false positive pixels, and TN_i is the number of true negative pixels.

2) Baseline: We employ the bounding box and the lowest value as negative point as auto-prompt for SAM and name it as "CDV prompt". To assess the quality enhancement from SAM in explaining visual concepts, we incorporate the raw heat maps with a threshold at 0.5, referred to as "threshold@0.5" for coarse semantic localization without using SAM. To compare the effectiveness of "CDV prompt" and human-constructed prompts, we use annotators' points on bird parts as positive prompt points of SAM. For all methods, images are scaled to 224×224 resolution. As SAM produce multiple masks, only the mask with the highest confidence score is chosen for the evaluation.

TABLE III

COMPARING DIFFERENT SEGMENTATION OF VISUAL CONCEPT USING *mIoU* ON CUB TESTSETS. WE INVESTIGATE THE DIFFERENCE ON THE SUBSET OF CORRECT AND INCORRECT PREDICTIONS. SAM(HUMAN) IS HUMAN ANNOTATED POINTS OF VISUAL CONCEPTS

Method	Correct (n=4560)	Incorrect (n=934)	All test data
Threshold@0.5	0.3630 ± 0.1148	0.2794 ± 0.1195	0.3374 ± 0.1224
SAM(human prompt)	0.3715 ± 0.1638	**0.3992** ± **0.2366**	0.3806 ± 0.1979
SAM(CDV prompt)	**0.6239** ± **0.1655**	0.1643 ± 0.2167	**0.4831** ± **0.2798**

3) Results: For quality comparison, Figure 9a shows randomly selected cases to visualize result of different methods, together with some hard cases shown in Figure 9b. For quantitative evaluation, Table III presents a comparison of IoU values across the three methods. According to above results, we have following observations and discussions:

- CDV's heat maps can roughly localize the bird's position but result in small holes after simply applying a 50% threshold. Thus, direct utilization of the thresholded heat map area for visual interpretation is insufficient.
- As shown in the fourth column of Figure 9a, the background is erroneously segmented as foreground in these cases for annotator-provided human prompt. Thus, direct utilization of positive prompts for SAM's cannot successfully interpret critical visual concept in image.
- The fifth column of Figure 9a presents the mask obtained by SAM following CDV auto-prompt. This mask exhibits a high consistency with the ground truth segmentation, indicating that SAM significantly improves visual semantic segmentation. This discrepancy can be attributed to CDV's inclusion of negative prompt points that make SAM pay less attention on background.

TABLE IV
ACCURACY ON VALIDATION SET WITH DIFFERENT CONCEPT NUMBERS

Dataset Type / Dataset Name	Num.	Natural	Semantic		Fine-grained		Specialized				Avg.
		CIFAR100	DTD	UCF101	CUB	Aircraft	EuroSAT	HAM10000	DR	Kera	
Label Free CBM	3	36.64%	59.22%	84.25%	54.15%	30.90%	63.41%	68.00%	53.31%	51.15%	55.67%
	5	45.09%	63.21%	86.56%	57.05%	34.26%	75.85%	69.60%	53.93%	49.51%	59.45%
	7	49.33%	65.25%	87.20%	58.30%	35.43%	84.11%	69.10%	53.24%	50.49%	61.38%
	10	52.33%	67.99%	89.46%	59.60%	36.42%	87.78%	71.50%	52.69%	48.52%	62.92%
	20	54.50%	67.73%	89.88%	60.20%	37.26%	90.78%	72.30%	56.62%	51.80%	64.56%
LaBo	3	78.85%	72.61%	96.47%	79.90%	57.55%	89.85%	71.30%	45.23%	52.65%	71.60%
	5	79.10%	75.27%	96.68%	80.65%	58.96%	93.15%	73.90%	47.10%	52.46%	73.03%
	7	79.71%	74.91%	97.10%	80.45%	60.34%	94.19%	75.40%	49.93%	53.79%	73.98%
	10	79.77%	75.35%	97.21%	81.00%	60.85%	94.74%	76.10%	50.55%	51.33%	74.10%
	20	80.03%	76.68%	97.42%	81.10%	61.45%	95.63%	79.60%	49.72%	53.41%	75.00%
Ours	3	78.37%	75.94%	92.05%	80.11%	59.53%	96.80%	83.12%	57.79%	70.69%	77.16%
	5	81.25%	79.02%	97.98%	82.62%	61.45%	96.79%	83.60%	57.90%	70.23%	78.98%
	7	81.68%	79.47%	98.12%	83.69%	62.18%	96.96%	83.58%	58.52%	71.15%	79.48%
	10	81.99%	79.43%	98.19%	83.97%	62.39%	**96.99%**	**83.78%**	58.54%	**71.48%**	79.64%
	20	**82.17%**	**79.79%**	**98.36%**	**84.26%**	**62.68%**	96.96%	83.72%	**58.58%**	71.61%	**79.79%**

TABLE V
ABLATION STUDY OF INITIALIZATION AND ADVERSARIAL TRAINING ON VIT-L-14

Method	Natural	Semantic		Fine-grained		Specialized				Avg.
	CIFAR100	DTD	UCF101	CUB	Aircraft	EuroSAT	HAM10000	DR	Kera	
DCBM(random)	75.76%	**78.36%**	86.30%	**78.28%**	**52.51%**	95.97%	79.98%	**53.19%**	67.47%	**74.20%**
DCBM(w/o ψ)	**75.62%**	77.62%	86.10%	77.24%	51.88%	**96.04%**	**81.03%**	52.90%	67.23%	73.96%
DCBM	75.37%	77.39%	85.60%	77.36%	50.25%	95.44%	80.74%	52.21%	**67.77%**	73.57%

4) Failure Cases Analysis: Figure 9b illustrates some instances of failure in visual localization. These failures can be attributed to various reasons, such as unclear foreground-background boundaries (1st row), complex visual content (3rd row), bias of SAM (2nd and 4th rows), or misplaced CDV prompts (5th row).

Table III presents a quantitative comparison of segmentation results across the three methods with *mIoU* values. On all test data, the mean IoU obtained using CDV prompts exceeds that of human prompts by 0.1025, suggesting the superiority of CDV-generated prompts. Using SAM with "CDV prompt" surpasses "Threshold@0.5" by 0.1457 in *mIoU*, highlighting SAM's substantial enhancement in the quality of explanations. As the quality of segmentation is closely related to the quality of the CDV prompt, we further report the *mIoU* on the subsets where DCBM predicts correctly and incorrectly respectively. Though "CDV prompt" significantly outperform others in correct predictions (n=4560), it is lower than "Threshold@0.5" by 0.1151 and than "human prompt" by 0.2349 when predicting erroneous ones (n=934). This discrepancy underscores the correlation between DCBM's image interpretation accuracy and classification accuracy. Inaccurate segmentation often indicates incorrect recognition of visual information by DCBM. Therefore, inaccurate segmentation results can help users identify instances where DCBM's classification is incorrect.

D. Ablation Study

1) Number of CDVs: To evaluate the performance of VLM-based CBMs with varying concept numbers, we assigned [3, 5, 7, 10, 20] concepts to each category. Table IV presents the accuracy of each model with an equal concept count. The model's performance generally improves with an increased number of concepts. Notably, our method maintains high performance even with a limited number of concepts and excels as the count grows, demonstrating its ability to achieve accurate classification without relying on an extensive concept inventory.

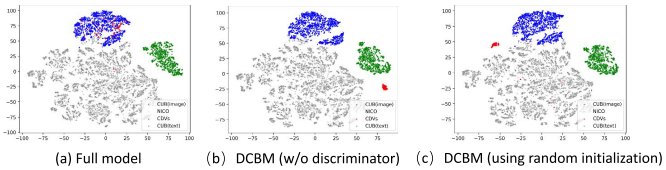


Fig. 10. t-SNE visualization of visual embedding, where blue points are embeddings of birds and green points are bird-related texts. The red points are the learned CDVs.

2) Discriminator and CDV Initialization: The discriminator ψ in Eq. 7 ensures the visual semantics of the learned concept vector. Adversarial training effectively minimizes the distance between distributions, guaranteeing that CDVs capture the true semantics of the training set. The initialization of CDV (e_i) impacts the model's performance and interpretability.

Our ablation study in Table V examines the impact of omitting initialization (DCBM random) and the discriminator (DCBM w.o ψ) on ViT-L-14 across various datasets. Interestingly, the performance of DCBM w.o ψ and DCBM random tends to surpass that of the full model, even matching the uninterpretable linear probe, suggesting a potential benefit from additional parameters.

Nonetheless, a slight performance reduction of 0.63% is an acceptable trade-off for enhanced interpretability. As shown in Figure 10, the full model's CDVs align with the features of bird images, whereas the CDVs from DCBM w.o ψ and DCBM(random) fail to correspond with vision and language, indicating a lack of the intended visual semantics.

VII. Conclusion

Existing VLM-based CBMs rely on pre-defined text concepts, which are subjective and ignored the influence modality gap. We address this through training CDVs to replace pre-defined text embeddings. Our proposed DCBM is able to achieve better performance than previous VLM-based CBM in both interpretability and accuracy, which is consistent with our hypothesis. As the previous methods are single-modal, for a fair comparison, we evaluate text interpretation and image interpretation separately. To the best of our knowledge, the quantitative evaluation of text interpretation does not have a universally established benchmark at present [62], [63], we conduct human study with multiple metrics, where more participants (n=27) were included compared to previous literature [8] (n=4). Although the current results are significant, more participants could certainly make the results more robust. Quantitative evaluation of text concept would be an essential future work, where a large-scale dataset is required to measure explainable artificial intelligence [63]. Another limitation is that our method relies on a multi-modal model that has been pre-trained with contrastive loss. Though we have addressed the performance decrease caused by modality gap, it still posed challenges in accurately mapping visual concepts to textual interpretation, especially in ambiguous or context-dependent scenarios. One of the possible reason is the concept associate bias in the pre-trained model [35]. Future research can consider the impact of other loss terms, such as additional image caption task in BLIP2 [33] and explore strategies to eliminate concept associate bias.

Another contribution of this research is the exploration of cross-modal interpretation using the proposed notions of concept, and the utilization of concept-sample distribution (CSD) to describe the relationship between concepts and samples. Through our investigations, we have discovered that CSD can be used to learn the same concept in multiple modalities by leveraging the contrastive loss. Exploiting this finding, we have successfully employed a modality inverter to predict the representation of concepts in the middle layer of the CLIP visual encoder. Furthermore, we obtain the location of interpreted visual concepts by constructing automatic prompts for Segment Anything Model (SAM) [11]. Experimental results show that the segmentation with automatic prompts is consistent with the ground-truth, surpassing the performance of human prompts. This improvement can be attributed to the ability of the auto-prompts to acquire both positive and negative annotations easily, while human prompts only have positive annotations. However, it should be noted that the parameters of SAM are not modified in this study, leaving the exploration of scenario-specific SAM variants as future work. In summary, this paper applies the concept quintuple notion and CSD theory to explore cross-modal interpretation by training CDV to mitigate the influence of modality gap in existing CBM. The proposed DCBM not only enhances classification performance but also provides rich interpretation information for various applications. Additionally, we present a preliminary application of SAM in the field of interpretable classification, indicating the potential of SAM in improving the interpretability of models.

References

[1] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019.

[2] M. Blum and L. Blum, "A theoretical computer science perspective on consciousness," *J. Artif. Intell. Consciousness*, vol. 8, no. 1, pp. 1–42, Mar. 2021.

[3] Y. Pan, "Structure analysis of crowd intelligence systems," *Engineering*, vol. 25, pp. 17–20, Jun. 2023.

[4] P. W. Koh et al., "Concept bottleneck models," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5338–5348.

[5] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[6] M. Yuksekgonul, M. Wang, and J. Zou, "Post-hoc concept bottleneck models," in *Proc. 11th Int. Conf. Learn. Represent.*, Sep. 2022, pp. 1–20.

[7] T. Oikarinen, S. Das, L. M. Nguyen, and T.-W. Weng, "Label-free concept bottleneck models," in *Proc. 11th Int. Conf. Learn. Represent.*, Sep. 2022, pp. 1–32.

[8] Y. Yang, A. Panagopoulou, S. Zhou, D. Jin, C. Callison-Burch, and M. Yatskar, "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 19187–19197.

[9] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun, "Pre-trained language models and their applications," *Engineering*, vol. 25, pp. 51–65, Jun. 2023.

[10] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 17612–17625.

[11] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2023, pp. 4015–4026.

[12] S. M. Frankland and J. D. Greene, "Concepts and compositionality: In search of the brain's language of thought," *Annu. Rev. Psychol.*, vol. 71, no. 1, pp. 273–303, Jan. 2020.

[13] Y. Jiang et al., "Artificial intelligence for retrosynthesis prediction," *Engineering*, vol. 25, pp. 32–50, Jun. 2023.

[14] L. Yuan and S.-C. Zhu, "Communicative learning: A unified learning formalism," *Engineering*, vol. 25, pp. 77–100, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095809923001339, doi: 10.1016/j.eng.2022.10.017.

[15] D. Marcos, S. Lobry, and D. Tuia, "Semantically interpretable activation maps: What-where-how explanations within CNNs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 4207–4215.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[17] B. Kim et al., "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 2668–2677.

[18] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *Proc. 36th Int. Conf. Mach. Learn.*, May 2019, pp. 2376–2384.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 1135–1144.

[20] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf

[21] F. Tuo, Q. Ruijie, W. Xiaohan, W. Wenguan, and Y. Yi, "Interpretable3D: An ad-hoc interpretable classifier for 3D point clouds," in *Proc. 38th AAAI Conf. Artif. Intell.*, vol. 32, 2024, pp. 1761–1769.

[22] M. Nauta, J. Schlötterer, M. van Keulen, and C. Seifert, "PIP-Net: Patch-based intuitive prototypes for interpretable image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2744–2753.

[23] Z. Fang et al., "Enabling collaborative clinical diagnosis of infectious keratitis by integrating expert knowledge and interpretable data-driven intelligence," 2024, *arXiv:2401.08695*.

[24] F. Kong, Y. Li, H. Nassif, T. Fiez, R. Henao, and S. Chakrabarti, "Neural insights for digital marketing content design," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2023, pp. 4320–4332.

[25] W. Wang, C. Han, T. Zhou, and D. Liu, "Visual recognition with deep nearest centroids," Mar. 2023, *arXiv:2209.07383*.

[26] C. Ma, B. Zhao, C. Chen, and C. Rudin, "This looks like those: Illuminating prototypical concepts using multiple visualizations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–24.

[27] J. Donnelly, A. J. Barnett, and C. Chen, "Deformable ProtoPNet: An interpretable image classifier using deformable prototypes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10265–10275.

[28] M. E. Zarlenga et al., "Concept embedding models: Beyond the accuracy-explainability trade-off," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. Curran Associates, 2022, pp. 21400–21413. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/867c06823281e506e8059f5c13a57f75-Paper-Conference.pdf

[29] A. Mahinpei, J. Clark, I. Lage, F. Doshi-Velez, and W. Pan, "Promises and pitfalls of black-box concept learning models," Jun. 2021, *arXiv:2106.13314*.

[30] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 26, 2024, doi: 10.1109/TPAMI.2024.3369699.

[31] A. Sain, A. K. Bhunia, P. N. Chowdhury, S. Koley, T. Xiang, and Y.-Z. Song, "CLIP for all things zero-shot sketch-based image retrieval, fine-grained or not," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 2765–2775.

[32] A. Q. Nichol et al., "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2022, pp. 16784–16804.

[33] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Machine Learn.*, 2023, pp. 19730–19742.

[34] H. Zhang et al., "GLIPv2: Unifying localization and vision-language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, Oct. 2022, pp. 36067–36080.

[35] Y. Yamada, Y. Tang, and I. Yildirim, "When are lemons purple? The concept association bias of CLIP," Dec. 2022, *arXiv:2212.12043*.

[36] S. Menon and C. Vondrick, "Visual classification via description from large language models," in *Proc. 11th Int. Conf. Learn. Represent.*, Sep. 2022, pp. 1–17.

[37] J. Gu et al., "A systematic survey of prompt engineering on vision-language foundation models," Jul. 2023, *arXiv:2307.12980*.

[38] J. Wang et al., "Review of large vision models and visual prompt engineering," *Meta-Radiol.*, vol. 1, no. 3, 2023, Art. no. 100047, doi: 10.1016/j.metrad.2023.100047.

[39] P. Gao et al., "CLIP-adapter: Better vision-language models with feature adapters," 2021, *arXiv:2110.04544*.

[40] R. Zhang et al., "Tip-adapter: Training-free CLIP-adapter for better vision-language modeling," 2021, *arXiv:2111.03930*.

[41] E. J. Hu et al., "LoRa: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.

[42] H. Yao, R. Zhang, and C. Xu, "Visual-language prompt tuning with knowledge-guided context optimization," 2023, *arXiv:2303.13283*.

[43] W. Ji, J. Li, Q. Bi, T. Liu, W. Li, and L. Cheng, "Segment anything is not always perfect: An investigation of SAM on different real-world applications," *Mach. Intell. Res.*, vol. 21, pp. 1–14, 2024. [Online]. Available: https://www.mi-research.net/en/article/doi/10.1007/s11633-023-1385-0

[44] T. Chen et al., "SAM-adapter: Adapting segment anything in underperformed scenes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2023, pp. 3367–3375.

[45] L. P. Osco et al., "The segment anything model (SAM) for remote sensing applications: From zero to one shot," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 124, Nov. 2023, Art. no. 103540.

[46] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: An experimental study," *Med. Image Anal.*, vol. 89, Oct. 2023, Art. no. 102918.

[47] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[48] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5583–5594.

[49] A. Krizhevsky and G. Hintonm, "Learning multiple layers of features from tiny images," Toronto, ON, Canada, 2009.

[50] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3606–3613.

[51] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[52] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200–2011 dataset," California Inst. Technol., 2011.

[53] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," 2013, *arXiv:1306.5151*.

[54] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.

[55] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, Aug. 2018.

[56] S. D. Karthik. (2019). *Aptos 2019 Blindness Detection*. Maggie. [Online]. Available: https://kaggle.com/competitions/aptos2019-blindness-detection

[57] Z. Fang, K. Kuang, Y. Lin, F. Wu, and Y.-F. Yao, "Concept-based explanation for fine-grained images and its application in infectious keratitis classification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 700–708.

[58] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.

[59] E. Wong, S. Santurkar, and A. Madry, "Leveraging sparse linear layers for debuggable deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11205–11216.

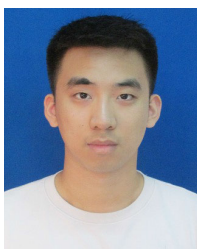
[60] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2019, Paper 832.

[61] M. Velmurugan, C. Ouyang, C. Moreira, and R. Sindhgatta, "Developing a fidelity evaluation approach for interpretable machine learning," Jun. 2021, *arXiv:2106.08492*.

[62] P. Q. Le, M. Nauta, V. B. Nguyen, S. Pathak, J. Schlötterer, and C. Seifert, "Benchmarking explainable AI: A survey on available toolkits and open challenges," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, Macau, SAR China, Aug. 2023, pp. 6665–6673.

[63] M. Nauta et al., "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI," *ACM Comput. Surv.*, vol. 55, no. 13, pp. 1–42, Jul. 2023.

Zhengqing Fang received the B.Sc. degree from the College of Computer Science, Zhejiang University, in 2019. He is currently pursuing the Ph.D. degree in ophthalmology and artificial intelligence with Zhejiang University School of Medicine. His research interests include explainable AI, medical image analysis, and data mining.



Zhouhang Yuan received the B.Sc. degree from the College of Computer Science, Zhejiang University, in 2022. He is currently pursuing the Ph.D. degree in ophthalmology and artificial intelligence with Zhejiang University School of Medicine. His research interests include artificial intelligence in medicine.

Ziyu Li is currently pursuing the Ph.D. degree in computer science and technology with the College of Computer Science and Technology, Zhejiang University, Hangzhou, China. His research interests include artificial intelligence in medicine.

Jingyuan Chen received the B.Eng. degree from Beihang University and the Ph.D. degree from the National University of Singapore (NUS). She is currently a ZJU100 Young Professor. Her research interests include AI for education, multimedia content analysis, and information retrieval.

Kun Kuang received the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, in 2019. He was a Visiting Scholar with the Prof. Susan Athey's Group, Stanford University. He is currently an Associate Professor with the College of Computer Science and Technology, Zhejiang University. He has published more than 100 papers in prestigious conferences and journals in data mining and machine learning, including *Cell Patterns*, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, ICML, NeurIPS, KDD, ICDE, WWW, SIGIR, and ACM Multimedia. His main research interests include causal inference, causality-inspired machine learning, and smart justice. He received the ACM SIGAI China Rising Star Award in 2022.

Yu-Feng Yao received the M.D. degree from Zhejiang Medical University in 1984, and the Ph.D. degree from Osaka University, Japan. From 1992 to 1997, he studied corneal diseases under the mentorship of Prof. Yasuo Tano and Prof. Yuichi Ohashi with Osaka University. He is currently a Professor and Ophthalmologist with the Department of Ophthalmology, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine. He is known for his expertise in corneal disease treatment and has developed innovative surgical techniques that are included in U.S. medical textbooks.

Fei Wu (Senior Member, IEEE) received the B.Sc. degree in computer science from Lanzhou University in 1996, the M.Sc. degree in computer science from the University of Macau in 1999, and the Ph.D. degree in computer science from Zhejiang University in 2002. He is currently a Qiushi Distinguished Professor with the College of Computer Science, Zhejiang University. He is also the Dean of the College of Computer Science and the Director of the Institute of Artificial Intelligence, Zhejiang University.