CoE: Chain-of-Explanation via Automatic Visual Concept Circuit Description and Polysemanticity Quantification

Wenlong Yu Qilong Wang* Chuang Liu Dong Li Qinghua Hu
Tianjin Key Lab of Machine Learning, College of Intelligence and Computing, Tianjin University
{ywl_95, qlwang, liuc_09, li_dong, huqinghua}@tju.edu.cn

Abstract

*Explainability is a critical factor influencing the wide deployment of deep vision models (DVMs). Concept-based post-hoc explanation methods can provide both global and local insights into model decisions. However, current methods in this field face challenges in that they are inflexible to automatically construct accurate and sufficient linguistic explanations for global concepts and local circuits. Particularly, the intrinsic polysemanticity in semantic Visual Concepts (VCs) impedes the interpretability of concepts and DVMs, which is underestimated severely. In this paper, we propose a **C**hain-**o**f-**E**xplanation (**CoE**) approach to address these issues. Specifically, CoE automates the decoding and description of VCs to construct global concept explanation datasets. Further, to alleviate the effect of polysemanticity on model explainability, we design a concept polysemanticity disentanglement and filtering mechanism to distinguish the most contextually relevant concept atoms. Besides, a Concept Polysemanticity Entropy (CPE), as a measure of model interpretability, is formulated to quantify the degree of concept uncertainty. The modeling of deterministic concepts is upgraded to uncertain concept atom distributions. Finally, CoE automatically enables linguistic local explanations of the decision-making process of DVMs by tracing the concept circuit. GPT-4o and human-based experiments demonstrate the effectiveness of CPE and the superiority of CoE, achieving an average absolute improvement of 36% in terms of explainability scores.*

1. Introduction

Deep learning-based vision models have demonstrated outstanding performance in various visual tasks, owing to the powerful learning and representation capabilities of deep neural networks (DNNs) [29, 50]. However, the extreme complexity of DNNs, while enhancing their performance, also limits their explainability [41]. Humans are often unaware of the reasons behind the model's conclusions. This lack of explainability significantly hinders the wide application of deep vision models (DVMs) in critical fields such as autonomous driving and medical diagnosis [7, 28].

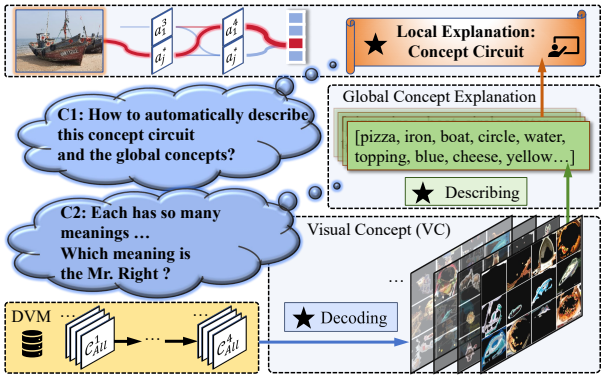


Figure 1. Illustration of the integrated process of concept-based post-hoc XAI methods [36], along with two challenges they face.

Post-hoc eXplainable Artificial Intelligence (XAI) can enhance the transparency of DVMs without causing degradations in performance [34, 43]. Research in this field can roughly be divided into local and global XAI methods [6]. The former explains either specific stages or the entire decision path with respect to a specific sample, identifying key activated regions or significant semantic concepts [3, 45, 46]. The latter deciphers the overall behavior of the DVM in a dataset by concept discovery or visualization techniques [22, 27, 52]. Among them, concept-based methods can explain DVMs from both global and local (glocal) perspectives [2, 4], assuming that each channel or neuron in a certain layer is responsible for recognizing essential semantic concepts. As shown in Fig. 1, they primarily perform one or two of the three operations: decoding Visual Concepts (VCs) of DVMs, manually constructing global concept explanation datasets, and providing local concept circuit explanations for individual samples [14, 16, 26].

Despite the significant progress made in concept-based explanation research, certain challenges remain, as illustrated in Fig. 1. First, the VCs are decoded by manually

*Corresponding author.

identifying and annotating the commonalities of a set of extracted image patches [2, 26]. These manual methods intuitively become ineffective when faced with new scenarios, new models, and new XAI methods, due to high labor costs, low efficiency, and the limited cognitive scope of individual annotators. Second, most methods make a naive assumption: each concept is monosemantic and user-friendly [33]. However, in practice, VCs encapsulate polysemantic information [23] (*e.g.*, an average of 13 concept atoms are coupled per channel in a ResNet152 model), presenting challenges for the model explanation. Particularly, polysemanticity can serve as a metric for evaluating a model's interpretability. A higher level of polysemanticity signifies a broader conceptual space, increasing the risks of selecting incorrect semantics and generating deviated explanations. Arbitrary concept selection or presenting them solely as images without descriptions yields suboptimal explanations [43]. Both visualizations and linguistic explanations become challenging for humans to comprehend, as it is unclear which semantic is actually represented or holds prominence [16]. Although polysemanticity is crucial to XAI, its systematic quantification and resolution remain unaddressed. All of these challenges collectively exacerbate the difficulty of concept-based explanations for DVMs.

To alleviate this dilemma, we propose a Chain-of-Explanation (CoE) approach via automatic concept decoding, disentanglement, filtering, description, and polysemanticity quantification. Specifically, CoE first proposes an Automatic Concept decoding and Description (ACD) method to construct global concept explanation datasets. It decodes every VC of key layers through arbitrary advanced XAI methods. A Large Vision Language Model (LVLM) is utilized to describe the commonalities of the extracted image patches since the linguistic explanation is more comprehensible than images [43]. As for the polysemantic concepts, we formulate a Concept Polysemanticity Disentanglement and Filtering (CPDF) mechanism to disentangle them into a set of orthogonal concept atoms, along with a Concept Polysemanticity Entropy (CPE) to quantify their degree of polysemanticity. In this paper, 13 distinct semantic directions are prompted, including low-level and high-level semantics, which is much more than previous work (*e.g.*, 5 in [56]). Within the CPDF, a semantic entailment model is applied to mitigate the impact of redundant semantics. After these operations, the modeling of deterministic concepts is upgraded to uncertain concept atom distributions. When elucidating the local decision-making process, a concept circuit method is utilized to form the basis of an explanation chain, with each node representing a hierarchical set of essential concepts. Relevant atoms and their interconnections are contextually filtered and aligned within the chain. Finally, we consolidate all conceptual information along the explanation chain to generate linguistic explanations similar to a Chain-of-Thought (CoT) pipeline [21, 51], leveraging a Large Language Model (LLM), as shown in Fig. 2. Qualitative and quantitative experiments validate the effectiveness and the superiority of the CoE approach. The contributions of this paper are summarized as follows.

- We propose a novel CoE approach to automatically and systematically provide concept-based glocal explanations, along with managing the problem of polysemanticity, in a more interpretable natural language format.
- A CPDF mechanism is formulated to handle the hard-to-interpret polysemantic concepts. To our knowledge, CPDF is a pioneer tailored to comprehensively disentangle and quantify polysemanticity. It disentangles them into a set of succinct concept atoms, followed by a filtering function to contextually find the suitable atom as the explanation node. A CPE score is defined to quantify the polysemanticity of concepts, which can be utilized as a metric of the interpretability of concepts or models.
- Both GPT-4o and human-based experiments validate the superiority of the CoE in automatically constructing global and local explanations and the effectiveness of CPE in quantifying concept polysemanticity.

2. Related Work

Research on XAI can be broadly categorized into ad-hoc model interpretability and post-hoc explanation studies [43, 54]. The former actively refine DVMs to improve interpretability while also necessitating consideration of a trade-off between performance and interpretability [8, 31, 37, 42, 55]. The latter explains DVMs passively without performance decline [5, 11, 22, 45, 46]. Among these post-hoc efforts, concept-based explanations can provide glocal insights into DVMs [36]. Literature [56] annotated every pixel in special datasets with only 5 kinds of concepts. The explanation is derived by analyzing the activation regions of each neuron. MILAN [26] decoded VCs by calculating the maximum mutual information between channels and regions, followed by manually summarizing commonalities into 3 sentences. However, defining concepts solely through activations is insufficient. CRP decoded VCs through the relevance value derived from the deep Taylor decomposition [35], showing more accurate results than the activation-based methods [2]. The commonalities of the VCs were also identified manually, and one of these commonalities was selected randomly. However, these methods encounter significant challenges, as discussed in Sec. 1. In contrast, this paper proposes an automated concept decoding and description method based on LVLM, which offers greater flexibility in providing linguistic explanations. The most suitable concepts for interpretation are contextually selected, showing superior results.

Circuits, as sub-graphs, are proposed to explain the key decision route of a model [3, 13, 17, 20, 48]. They identify

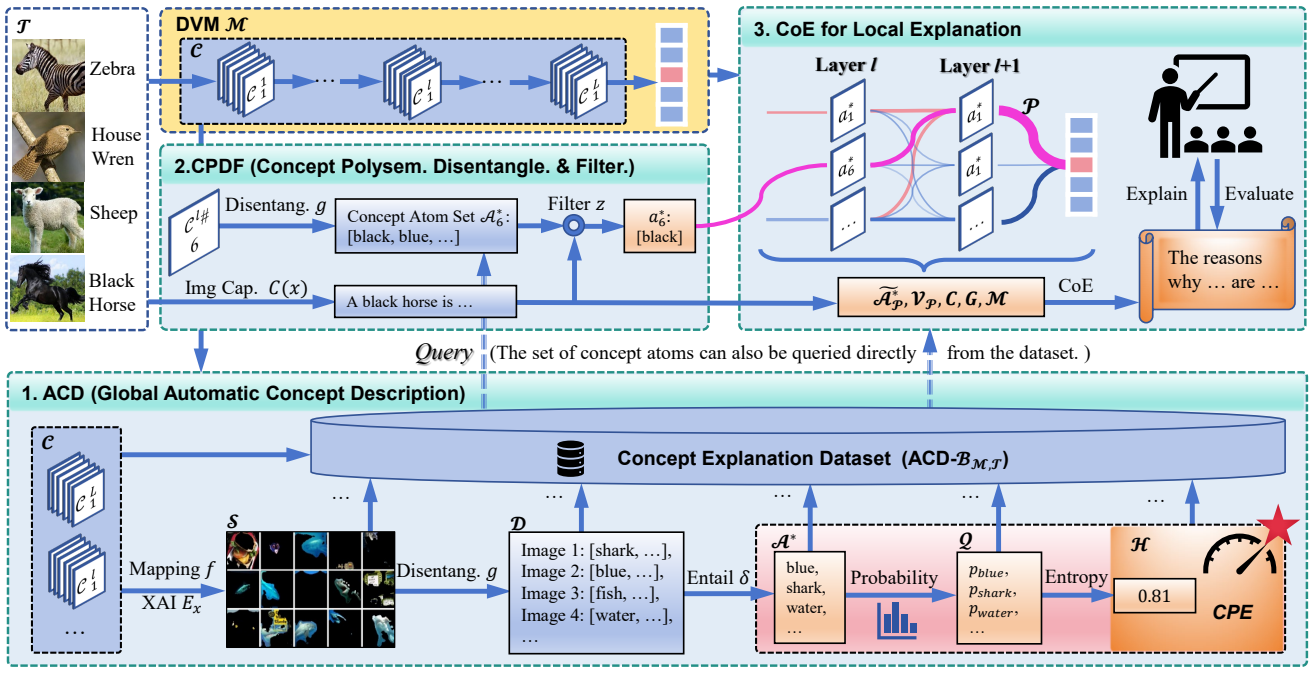


Figure 2. Outline of the Chain-of-Explanation. It consists of three steps (i.e., ACD, CPDF, and local explanation). Each part utilizes a set of channels $\mathcal{C}$ from $\mathcal{M}$. In the concept circuit, $\mathcal{P}$, lines of different depths and colors represent the magnitude of the relevance.

the highest-activated concepts and then explain the model's decision by linking all concepts across the circuit. However, its interpretability for DVMs is limited due to the lack of a cohesive, language-based explanatory output. Additionally, the polysemanticity of VCs significantly hinders the accurate explanation [18, 33], and research in this area remains limited. Some methods attempted to indirectly mitigate its impact by predefining principal vectors or subspaces [20, 27, 38]. The lack of flexible and accurate concept-level explanations also restricts their adaptability. Literature [33] analyzed the polysemanticity from an information compression perspective. Pure [16] alleviated the polysemanticity by clustering and decoupling the circuit route. These methods, however, lack the ability to simultaneously and comprehensively quantify, disentangle, and describe polysemantic VCs. Their applicability is limited—challenges that this paper seeks to address, as discussed in Sec. 3.

3. Chain-of-Explanation Approach

In this paper, a CoE approach is proposed to automatically explain DVMs from both global and local perspectives. It breaks down the complex reasoning path into a chain of concept nodes, within which polysemanticity is effectively managed. As shown in Fig. 2, CoE is composed of three steps. **Step 1:** ACD is proposed to automatically construct global concept descriptive explanation datasets of a given DVM. **Step 2:** With the assistance of an LVLM, a CPDF mechanism is designed to disentangle each original VC into a set of concept atoms and extract the most suitable atom according to the context, making CoE more comprehensible and flexible. Besides, a CPE score is formulated accordingly to quantify the degree of polysemanticity of concepts. **Step 3:** A local explanation chain in a manner akin to a CoT for the decision-making progress of the DVM is established based on the dataset, CPDF, concept circuit, and LLMs.

3.1. Automatic Concept Decoding and Description

ACD method is first formulated to automatically and effectively construct databases of global concept descriptions. Specifically, given an image dataset $\mathcal{T}$ and a DVM $\mathcal{M}$ requiring evaluation and explanation, we establish an injective function f between channel set $\mathcal{C}$ and visual semantic concept set $\mathcal{S}$ by an XAI method E_x, $f : \mathcal{C} \mapsto \mathcal{S}$. Without loss of generality, for j_{th} channel c_j^l in layer l, E_x extracts a set of image patches within $\mathcal{T}$ that most activate this channel, named VC $s_j^l = \{I_n\}_N$, where N is the number of image patches. Each I_n highlights certain regions of the original image, representing the focused concept, as exemplified in the bottom left part of Fig. 2.

A describing function r is then defined to decipher $\{I_n\}_N$ into a set of language descriptions:

$$\mathcal{D} = r(\{I_n\}_N) = \{d_m\}_M\,, \tag{1}$$

where M is the number of language descriptions for the current channel. Traditionally, r is implemented by manually and subjectively identifying commonalities within $\{I_n\}_N$ [56], while ACD utilizes a powerful LVLM as the describer

with well-engineered prompts. After constructing $\mathcal{S}$, $\mathcal{D}$ and the mapping function f, a concept description dataset $\mathcal{B}_{\mathcal{M},\mathcal{T}} = \{\mathcal{C}, \mathcal{S}, \mathcal{D}, f\}$ is automatically constructed, providing global linguistic explanations for all VCs of $\mathcal{M}$.

3.2. Managing Polysemantic Concept

VCs of each node within a DVM exhibit polysemanticity, introducing certain deviations in the explanation framework, as shown in Fig. 2. Higher polysemanticity indicates that the current VC aggregates more semantics, making the intended representation more ambiguous. Such semantic uncertainty impedes human comprehension of the concept, particularly within the CoE, where various concepts must be understood collectively. Quantifying and solving this dilemma is difficult and has been underestimated by previous studies. Under this circumstance, we propose a CPDF mechanism along with a CPE score to mitigate and quantify its detriments to the CoE simultaneously.

3.2.1 Polysemantic Concept Decoupling and Filtering

Considering that a polysemous word can be understood as an aggregation of various distinct meanings, we define a function g to disentangle the ambiguous VC into a set of easily understandable linguistic concept atoms:

$$\mathcal{A} = \left\{a^1, \cdots, a^i, \cdots, a^Q\right\} = g(s), \tag{2}$$

where Q is the number of the concept atoms. We posit that each atom represents a view of the current VC s, which will be activated upon the appearance of the corresponding semantic in the input. Furthermore, given that concept atoms should accurately represent the commonalities in $\{I_n\}_N$ and recognizing that this naive disentanglement function (Eq. 2) does not provide a foundation for quantifying polysemanticity, we refine the function g to disentangle each image patch I_n into a set of atoms:

$$\begin{aligned} \mathcal{A}_n = \left\{a_n^1, \cdots, a_n^i, \cdots, a_n^Q\right\} = g(I_n, s), \\ \mathcal{D} = \{\mathcal{A}_n\}_N. \end{aligned} \tag{3}$$

We instantiate g by a powerful LVLM (*e.g.*, GPT-4o [39]), considering that this is quite an open and complex problem. g prioritizes disentangling and describing commonalities within the VC while analyzing each individual image. Various commonalities among any certain number of image patches should be summarized and described by precise terminologies. These common semantics within an open scenario are as diverse as they are abundant, unfolding in countless forms and nuances. Besides, it should generate specific atoms when any commonalities identified above do not appear in the current I_n. Therefore, the art of prompt engineering `prompt-com` becomes paramount. It should provide rules for recognizing commonalities, possible conceptual directions, and output format control. In this paper, we prompt 13 high-level and low-level semantical directions that cover the normal concepts encountered in daily life, including object category, scene, object part, color, texture, material, position, transparency, brightness, shape, size, edges, and their relationships. Additionally, the CoT techniques [21, 51] are introduced. Details of all sophisticated prompts are shown in Sec. S2 of the Appendix.

All $\mathcal{A}_n$ are then aggregated into a description atom set $\mathcal{A}$ with all duplicate atoms removed:

$$\begin{aligned} \mathcal{A} &= \{\mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots \cup \mathcal{A}_N\} \\ &= \left\{a^1, \cdots, a^i, \cdots, a^P\right\}, \end{aligned} \tag{4}$$

where P is the number of concept atoms after deleting the repeating atoms. Each element in $\mathcal{A}$ represents a disentangled monosemantic concept and can be utilized as a node of the CoE. However, some atoms are semantically equivalent even if they have variant descriptions [19]. In order to remove semantically redundant atoms, we introduce an entailment function δ to detect the relations between atoms:

$$\{1, 0, -1\} = \delta(a^{i1}, a^{i2}),\ \forall a^{i1}, a^{i2} \in \mathcal{A}, \tag{5}$$

where the responses 1,0,-1 represent entailment, neutral, or contradiction of the candidates, respectively. We instantiate δ by an NLI model (*e.g.*, DeBERTa model [25]) because the atom candidates are relatively short and simple that an NLI model is good enough. Following [19], a bidirectional entailment check is applied, meaning that one of the candidates will be removed when the response is entailment or neutral. The first atom is designated as the monosemantic representative of the semantic equivalence set due to the transitivity between them. The final concept set $\mathcal{A}^*$ only contains heterogeneous atoms that describe the current VC:

$$\mathcal{A}^* = \left\{a^1, \cdots, a^i, \cdots, a^{P^*}\right\}. \tag{6}$$

After disentangling the concept atoms, we define a filtering function z to select an atom a^* as the most suitable explanation of current VC with respect to the input x:

$$a_s^*(x) = z(\mathcal{A}_s^*, C(x)). \tag{7}$$

Some research ignores this step or selects the concept randomly with an implicit assumption that all concepts in $\mathcal{A}^*$ are semantically equivalent [14, 26]. In contrast, we filter the most suitable atom a^* based on the context of the CoE. In particular, z is instantiated by an LLM. $C(x)$ is an image captioning function utilized for providing contextual information. It can be instantiated by a relatively small LVLM [9, 10] or a professional model [30, 49].

3.2.2 Quantification of Concept Polysemanticity

The polysemanticity of concepts can serve as a metric for the model's interpretability. Lower polysemanticity of concepts indicates reduced semantic uncertainty with a more

singular form of activation, leading to greater interpretability. In this paper, we endeavor to quantify polysemanticity from three levels of granularity (i.e., conceptual channels, layers, and models) by defining a CPE score. The larger the CPE, the more pronounced the polysemanticity, which leads to decreased interpretability.

In particular, the disentanglement function g in Eq. 3 enables us to count the frequency of each atom in $\mathcal{A}^*$. Given two fixed numbers Q and N, we calculate the probability of the i_{th} atom within j_{th} concept:

$$p_i = \frac{Num_i}{Q \times N + Pad}. \tag{8}$$

Then, the CPE of j_{th} concept can be formulated as

$$H_j^* = \frac{-\sum_{i=1}^{P^*+Pad} p_i \log p_i}{\log(P^* + Pad)}. \tag{9}$$

The denominator normalizes the entropy value to a range between 0 and 1, considering that the size of P^* varies for different VCs. Besides, the quantity of $\mathcal{A}^*$ is set to be at least equal to N due to some $\mathcal{D}_s$ having a small number of semantical equivalent atoms with equal probabilities. If there are insufficient unique atoms, we supplement the corresponding number of atoms, with each supplementary atom set to a number of 1. The atomic objects are expanded, and the denominator is increased by the number of supplemented atoms (a total amount of Pad). Please refer to Sec. S3.1 of the Appendix for details. In addition, two averaged CPEs are formulated to enable comparisons of polysemanticity between specific layers or models:

$$H_l^* = \frac{1}{d_l} \sum^{d_l} H_j^*, \tag{10}$$

$$H_{\mathcal{M}}^* = \frac{1}{L} \sum^{L} H_l^*, \tag{11}$$

where d_l is the dimension of the channel and L is the number of layers. Finally, to accommodate the diverse needs of researchers within the community, the automatically constructed concept explanation dataset is extended as ACD-$\mathcal{B}_{\mathcal{M},\mathcal{T}} = \{\mathcal{C}, \mathcal{S}, \mathcal{D}, \mathcal{A}^*, \mathcal{Q}, \mathcal{H}, f\}$, where $\mathcal{Q}$ represents the probability set of each atom and $\mathcal{H}$ is the set of CPE.

3.3. Construction of Explanation Chain

Concept circuits in DVMs can identify the most activated units in each layer for a specific input [3]. When aggregating these units across all key layers in a bottom-up manner, a corresponding maximum activation path $\mathcal{P}$ is generated, establishing the foundation for the explanation chain. Each node in this path serves as the explanation for the model's decision at that layer. Traditional concept circuit methods for DVMs identify the activation path without providing thorough overarching natural language-based explanations. Inspired by the CoT explanation technique in LLMs [12], CoE proposes a local explanation chain to interpret the decision-making process of DVMs in the form of natural language descriptions based on ACD-$\mathcal{B}_{\mathcal{M},\mathcal{T}}$, CPDF, concept circuit method, and LLM, as shown in Fig. 2.

Specifically, some user-predefined layers requiring explanations or key layers facilitating CoE (*e.g.*, output layers of 4 stages of a DVM) are listed in a bottom-up manner. Given an image x and layer l, an advanced XAI method E_x computes the maximization normalized activation or relevance values for all concepts. Each value represents the importance of the current concept to the prediction $\mathcal{M}(x)$. CoE acknowledges top k^l relevant concepts $\mathcal{C}^l(x) = \{c_1^l, \cdots, c_{k^l}^l\}$ and their values $\mathcal{V}^l(x) = \{v_1^l, \cdots, v_{k^l}^l\}$, determined by the α quantile of relevance values across all concepts (i.e., $k^l = |\{v_j^l | v_j^l > \alpha \max(v^l), \forall j \in d^l\}|$, where $|\mathcal{X}|$ computes the number of elements). After that, the descriptive concept atoms of the current layer $\widetilde{\mathcal{A}}_{k^l}^{*,l}(x) = \{a_1^{*l}, \cdots, a_{k^l}^{*l}\}$ are constructed based on ACD-$\mathcal{B}_{\mathcal{M},\mathcal{T}}$ and the filter Eq. 7, serving as a node of the explanation chain.

The final CoE for the local decision-making explanation, abstracting and synthesizing all top activated concept atoms along the chain, is formulated as

$$CoE_{\mathcal{M}}(x) = e(\widetilde{\mathcal{A}}_{\mathcal{P}}^*(x), \mathcal{V}_{\mathcal{P}}(x), C(x), G(x), \mathcal{M}(x)), \tag{12}$$

where e formulates a synthesizing and describing function with inputs obtained from the above sections, and $G(x)$ is the label of x. We instantiate $CoE_{\mathcal{M}}(x)$ through a powerful LLM (*e.g.*, GPT-4 [1]), with prompt `prompt-coe` well-designed by combining CoT and few-shot prompting [44].

Additionally, these local explanations are evaluated by a powerful LVLM and human judges based on three explainability metrics, each utilizing a three-tier scoring system.

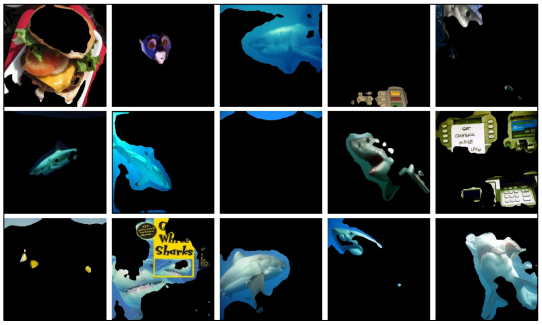
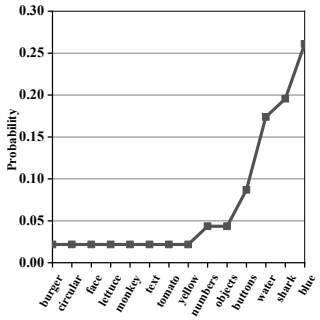
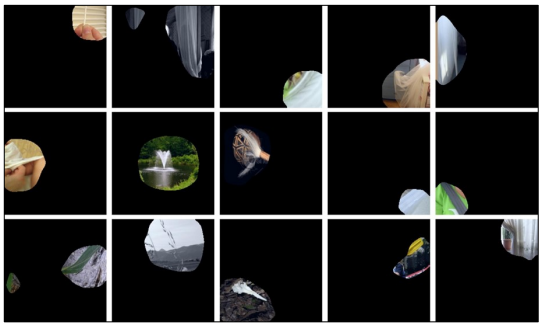
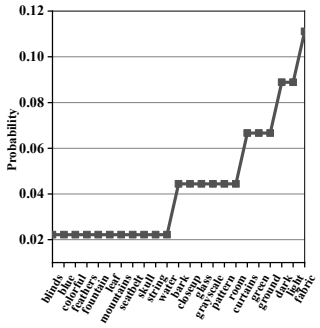
4. Experiments

We evaluate the proposed CoE approach on two types of DVMs (i.e., ResNet [24] and CLIP [40]). The relevance-based XAI method CRP is adopted as the primary approach for E_x in this paper, owing to its demonstrated superiority in terms of fidelity and reliability [2]. Please refer to the Appendix for more details on the implementation and results. Codes are available at https://github.com/YuWLong666/CoE.

4.1. Concept Explanation Dataset Construction

In the CoE, we first automatically construct global concept description databases ACD-$\mathcal{B}_{\mathcal{M},\mathcal{T}}$ for various DVMs. Taking a ResNet152 model as an example, 4 output layers of 4 stages (with 256, 512, 1024, and 2048 channels $\mathcal{C}$, respectively) are selected as the key layer set. CRP is applied as E_x to extract the VC set $\mathcal{S}$. Following [2], each VC is represented by 15 (i.e., N) image patches. The mapping function

Table 1. Examples of VCs and their disentangled concept atoms, along with their concept probability distributions and CPE scores. The first row shows concept 75 of Layer 4 extracted by CRP. The second row shows concept 697 of Layer 4 extracted by MILAN.

Visual Concept and Disentangled Concept Atoms	Concept Distribution	CPE
 Image 1: [burger, lettuce, tomato], Image 2: [blue, monkey, face], Image 3: [shark, blue, water], Image 4: [keypad, buttons, numbers], Image 5: [shark, blue, water], Image 6: [shark, blue, water], Image 7: [shark, blue, water], Image 8: [blue, circular, shape], Image 9: [shark, blue, water], Image 10: [keypad, buttons, numbers], Image 11: [blue, yellow, objects], Image 12: [shark, blue, text], Image 13: [shark, blue, water], Image 14: [shark, blue, water], Image 15: [shark, blue, water]		**0.81**
 Image 1: [hand, blinds, string], Image 2: [curtains, room, grayscale], Image 3: [fabric, green, closeup], Image 4: [curtains, room, light], Image 5: [reflection, glass, light], Image 6: [hand, fabric, texture], Image 7: [water, fountain, greenery], Image 8: [feathers, pattern, dark], Image 9: [fabric, blue, closeup], Image 10: [fabric, green, seatbelt], Image 11: [leaf, bark, texture], Image 12: [nature, mountain, grayscale], Image 13: [skull, ground, dark], Image 14: [fabric, colorful, pattern], Image 15: [curtains, window, light]		**0.95**

f from channels to VCs is built. We then employ a GPT-4o-2024-08-06 model as the function g to generate the description set $\mathcal{D}$ (i.e., Eq. 3). The disentangled and clustered atom set $\mathcal{A}^*$ (processed by a DeBERTa model), the probability set $\mathcal{Q}$, and the CPE set $\mathcal{H}$ are computed and gathered accordingly. Overall, ACD-$\mathcal{B}_{\mathcal{M},\mathcal{T}}$ containing 3,840 entries is constructed, with each entry comprising a channel, a CPE score, and 15 VC representations—each consisting of 3 concept atoms (i.e., Q)—and multiple orthogonal linguistic atoms, along with their associated probabilities.

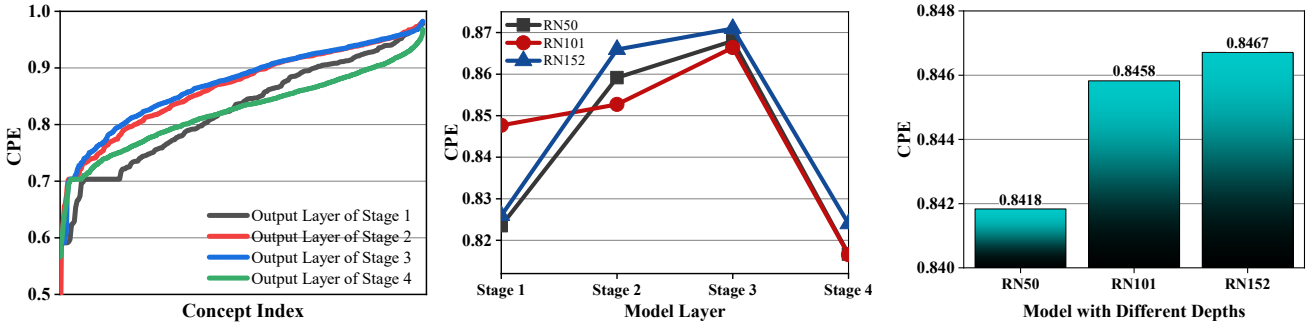
4.2. Disentanglement of Polysemantic Concept

CPDF automatically disentangles and describes the polysemantic VCs into a set of linguistic concept atoms, $\mathcal{D}$. As exemplified by Table 1, two randomly selected VCs are extracted by two advanced XAI methods (CRP and MILAN). They are disentangled and described into atoms, exhibiting different degrees of polysemanticity. We observe that the commonalities within the VCs are almost entirely identified, and most of the atoms describe the commonalities of subsets of image patches. In the first row (CPE=0.81), g successfully identifies all visually discernible commonalities, such as shark and water. The atoms exhibit high consistency, and fewer but more frequently occurring common semantics are identified. Only a few atoms, such as the burger, are not commonalities, as their corresponding images lack sufficient common semantics. The atoms are also represented accurately. For instance, the atom sets for each image featuring the shark semantic all include the word shark. In the second row (CPE=0.95), our method remains effective in disentangling and describing VC with a higher level of polysemanticity. g identifies a greater variety of atoms. These results demonstrate the effectiveness of the disentanglement and description operation in the CPDF.

4.3. Concept Distribution and CPE

CPE can quantify the polysemanticity and serve as an indicator of a DVM's interpretability. It begins by calculating the probabilities of each orthogonal atom. As the middle column of Table 1 shows, the distribution provides statistics on the extent to which different atoms are associated with each concept. For instance, concept 75 primarily focuses on blue with a probability of 0.27 while also identifying yellow with a smaller probability. The semantics between atoms are non-overlapping. This probabilistic approach enhances explainability by capturing multiple semantics with their occurring probabilities, upgrading the modeling of deterministic concepts to uncertain atom distributions.

CPE quantifies polysemanticity from 3 levels of granularity. As shown in Fig. 3a, concepts exhibit varying degrees of polysemanticity, with considerable fluctuations (ranging from a maximum of 1 to a minimum of 0.5). Most of the concepts demonstrate a noticeable level of polysemanticity. When analyzing the polysemanticity of different



(a) Sorted CPE along the channel dimension. (b) Averaged CPE on 4 key layers. (c) Averaged CPE on 3 models.

Figure 3. Illustrations of CPE scores at 3 levels of granularity. In (a), the channel indices of each layer are max normalized to 1.

Table 2. The consistency of the recognition of concept polysemanticity between the CPE metric and human scorers.

	Consistency
CPE Metric and Human Scorers	75 %

layers, as shown in Fig. 3b, the averaged CPE peaks in the third stage, while the shallowest and deepest layers exhibit lower values. Typically, deeper layers capture abstract concepts, whereas the shallower layers extract elementary features like textures [53, 56]. DVMs show better monosemanticity when identifying elementary and abstract concepts, while more ambiguous concepts are captured in the intermediate layers. This experiment validates the theoretical and experimental claims of previous work [14, 32]. In Fig. 3c, we observe that as the number of layers increases, the CPE of the model shows a rising trend. This result indicates a direct proportional relationship between the polysemanticity and the complexity of DVMs, which is that the more complex a DVM becomes, the harder it is to interpret.

The CPE is evaluated in comparison with human assessments. As shown in Table 1, the CPE is consistent with the degree of polysemanticity from the human intuitive observations (the polysemanticity of the first row is smaller than that of the second row). Similarly, we randomly sample 300 pairs of VCs, each containing two VCs with varying levels of polysemanticity. Comparisons are drawn from these 300 pairs by both human evaluators and the CPE metric. As shown in Table 2, results from human assessments and CPE achieve a 75% agreement rate, demonstrating the effectiveness of CPE. Other results are shown in Sec. S4 of the Appendix. All of these results reveal that CPE is effective in quantifying the polysemanticity of concepts and DVMs.

4.4. CoE for Local Explanations

CoE generates local explanations akin to CoT by automatically describing the concept circuit. We set $\alpha = 0.001$ to acknowledge the top relevant VCs, resulting in an average of 3 key VCs per layer. As exemplified in Table 3, given an input image, CoE first outputs whether the prediction of the DVM is correct and then gives reasons behind this prediction according to the information along the concept circuits. CoE employs an LLM as an engine to infer the logical relationships among concepts, as well as between concepts and the context. It synthesizes all evidence to generate an inductive commentary on the predictions. The explanation chain is presented in a linguistic format, enhancing human understanding. For instance, CoE explains why the DVM incorrectly predicted a tench as a mosquito net: the top activation of mosquito-net-related concepts, rather than those related to fish, misled the decision. Additional instances are shown in Sec. S5 of the Appendix. Overall, CoE effectively captures and elucidates the core decision-making evidence for both correct and incorrect predictions of DVMs.

4.5. Quantitative Evaluation via GPT-4o & Human

We employ GPT-4o and human evaluators to assess the local explanations for 500 and 100 samples randomly selected from the ImageNet Validation dataset [15], respectively. The GPT-4o evaluation is achieved through a carefully designed prompt `prompt-coe-eval`. Explanations generated by three methods are rated according to three explainability metrics: Accuracy, Completeness, and User Interpretability [47]. Each is assigned a three-tier scoring system (2, 1, and 0), with specific criteria defined for each tier, resulting in a total explanation score of 6. Please refer to Sec. S2 and S5 of the Appendix for details.

As shown in Table 4, our CoE approach attains peak performance across all three metrics, achieving total scores of 5.06 and 5.04 in GPT-4o and human evaluations, respectively. The lowest scores (3.14 and 2.62) are obtained on the MILAN ANNOTATION dataset without the ACD and CPDF methods. This dataset was manually annotated, resulting in inaccuracy and insufficiency within concept explanations. Scores increase by 0.11 and 0.72, respectively, with the application of our proposed ACD-$\mathcal{B}_{\mathcal{M},\mathcal{T}}$ dataset. The ACD method not only enables the automated disentanglement and description of concept atoms but also enhances their precision and comprehensibility. When the CoE em-

Table 3. Cases of local explanations of CoE. The first row shows the explanation of a correct prediction, and the second row is reversed.

 label: black-and-tan coonhound predict: black-and-tan coonhound	The model outputs a correct result: black-and-tan coonhound. Specifically, in the shallowest layer, layer 1, channel 88 with a relevance value of 1.0 describes the concept curved, which may relate to the curved shapes observed in parts of the dog such as its tail or ears. Moving deeper, in layer 2, channel 506 with a relevance value of 0.85 and channel 391 with a relevance value of 0.75 describe the concept of animal, indicating the model has identified an animal in the image, and channel 229 with a relevance value of 0.77 emphasizes the color black, aligning with the black and tan colors of the coonhound. Further, layer 2 also identifies the presence of grass several times (channels 94 and 103), which corresponds to the grassy field described in the image caption. In layer 3, various channels, such as 985, 174, 422, and 13, with high relevance scores, from 0.72 to 1.0, repeatedly identify the color green and the concept grass, firmly establishing the setting of the image. Additionally, channels 11, 787, and 591 in layer 3 enhance the certainty by repeatedly identifying animal and creature, confirming that the subject is an animal. Finally, layer 4 decisively contributes with channels 1354 and 1661, both signifying dog with high relevance values (1.0 and 0.72 respectively), which correctly identifies the animal in the image as a dog. Given that the model's structured concepts strongly match the description and characteristics noted in the caption and the ground truth label, it is clear why the model outputs the correct result of black-and-tan coonhound.
 label: tench predict: mosquito net	The model outputs an incorrect result mosquito net instead of the correct label tench. Specifically, in the shallowest layer 1 of the model, channel 88 with a relevance value of 1.0 describes the concept of smooth, channel 61 with a relevance value of 0.81 describes the concept of green, and channel 111 with a relevance value of 0.72 describes the concept of brown. In the deeper layer 2 of the model, channel 94 with a relevance value of 1.0 describes the concept of water, channel 229 with a relevance value of 0.84 describes the concept of fish, channel 103 with a relevance value of 0.81 also describes the concept of fish, and channel 467 with a relevance value of 0.73 describes the concept of mesh. At layer 3, channel 985 with a relevance value of 1.0 further describes the concept of fish, while channels 894, 122, and 715 with relevance values of 0.83, 0.73, and 0.72, respectively, describe the concept of green, possibly relating to the color of the net. In the deepest layer 4, channel 697 with a relevance value of 1.0 describes the concept of net, channel 1878 with a relevance value of 0.76 describes the concept of baby, and channel 321 with a relevance value of 0.76 represents the concept of hexagon, which could misconstrue the net structure. The prevalence of concepts related to fish within the decision path suggests a relationship to the tench, but the final prediction of mosquito net was likely influenced by the dominance of net-related concepts. Thus, the model was misled by the presence of a net, resulting in the incorrect classification of mosquito net.

Table 4. Scores of CoE for the local explanation on three methods evaluated by GPT-4o and human scorers. The first method utilizes MILAN ANNOTATION as the concept explanation dataset, while the rest two work on the ACD-$\mathcal{B}_{\mathcal{M},\mathcal{T}}$ datasets. **Each criterion has a maximum score of 2 points. The total explanation score is 6 points.** Higher scores indicate better explainability.

	Method	Accuracy	Completeness	User Interpret.	Total Explanation
GPT-4o	MILAN A.[26] + Description (Baseline)	1.03	1.07	1.04	3.14
	CoE w/o Filtering	1.07	1.10	1.08	3.25
	CoE (ours)	**1.69**	**1.70**	**1.67**	**5.06**
Human	MILAN A.[26] + Description (Baseline)	0.87	0.83	0.92	2.62
	CoE w/o Filtering	1.10	1.16	1.07	3.34
	CoE (ours)	**1.73**	**1.72**	**1.59**	**5.04**

ploys the full CPDF mechanism, the explanation score rises by 1.92 and 2.42 points (the average absolute improvement is 36%), respectively, with the filtering step contributing 1.81 and 1.70 points. This result suggests that by taking polysemanticity and contextual filtering of concepts into account, CoE effectively guides the LLM in producing better linguistic explanations, substantially enhancing their accuracy, completeness, and interpretability for users. The results of the human evaluation are consistent with the GPT-4o results. All of these results strongly affirm the superiority of the proposed CoE approach.

5. Conclusion

In this paper, we propose a CoE approach to tackle the challenges of the inflexibility to automatically construct concept explanation datasets, insufficient linguistic explanations, and the weakness in managing concept polysemanticity of concept-based XAI methods. CoE automates the decoding and description of concepts via an LVLM and constructs global concept explanation datasets. A CPDF mechanism is designed to identify the most contextually relevant concept atoms, advancing the modeling of deterministic concepts to uncertain concept atom distributions. The CPE is proposed to quantify the polysemanticity. Finally, CoE enables local explanations of a DVM's decision-making process, represented through natural language. The effectiveness and superiority of CPE and CoE are demonstrated. In summary, CoE establishes a systematical framework for explaining DVMs, and we believe that the findings presented in this paper warrant further exploration.

Acknowledgment

This research is supported in part by the National Natural Science Foundation of China under Grants U23B2049, 62276186, and 61925602.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[2] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019, 2023. 1, 2, 5

[3] Rajaram Achyuta, Chowdhury Neil, Torralba Antonio, Andreas Jacob, and Schwettmann Sarah. Automatic discovery of visual circuits. In *NeurIPS Workshop on Attributing Model Behavior at Scale*, 2023. 1, 2, 5

[4] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10, 2015. 1

[5] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020. 2

[6] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024. 1

[7] Subrato Bharati, M Rubaiyat Hossain Mondal, and Prajoy Podder. A review on explainable artificial intelligence for healthcare: why, how, and when? *IEEE Transactions on Artificial Intelligence*, 2023. 1

[8] Moritz Böhle, Navdeeppal Singh, Mario Fritz, and Bernt Schiele. B-cos alignment for inherently interpretable cnns and vision transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4504–4518, 2024. 2

[9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 4

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 4

[11] Ming-Ming Cheng, Peng-Tao Jiang, Ling-Hao Han, Liang Wang, and Philip Torr. Deeply explain cnn via hierarchical decomposition. *International Journal of Computer Vision*, 131(5):1091–1105, 2023. 2

[12] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203, 2024. 5

[13] Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36: 16318–16352, 2023. 2

[14] Meghal Dani, Isabel Rio-Torto, Stephan Alaniz, and Zeynep Akata. Devil: Decoding vision features into language. In *DAGM German Conference on Pattern Recognition*, pages 363–377. Springer, 2023. 1, 4, 7

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

[16] Maximilian Dreyer, Erblina Purelku, Johanna Vielhaben, Wojciech Samek, and Sebastian Lapuschkin. Pure: Turning polysemantic neurons into pure features by identifying relevant circuits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 8212–8217, 2024. 1, 2, 3

[17] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021. 2

[18] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022. 3

[19] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. 4

[20] Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[21] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4

[22] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019. 1, 2

[23] Dan W Grupe and Jack B Nitschke. Uncertainty and anticipation in anxiety: an integrated neurobiological and psy-

chological perspective. *Nature Reviews Neuroscience*, 14(7): 488–501, 2013. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[25] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2020. 4

[26] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep features. In *International Conference on Learning Representations*, 2022. 1, 2, 4, 8

[27] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 1, 3

[28] Anton Kuznietsov, Balint Gyevnar, Cheng Wang, Steven Peters, and Stefano V. Albrecht. Explainable ai for safe and trustworthy autonomous driving: A systematic review. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–23, 2024. 1

[29] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 4

[31] Haomiao Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. What is a tabby? interpretable model decisions by learning attribute-based classification criteria. *IEEE transactions on pattern analysis and machine intelligence*, 43(5): 1791–1807, 2019. 2

[32] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. 7

[33] Simon C Marshall and Jan H Kirchner. Understanding polysemanticity in neural networks through coding theory. *arXiv preprint arXiv:2401.17975*, 2024. 2, 3

[34] Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pages 1–66, 2022. 1

[35] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222, 2017. 2

[36] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019. 1, 2

[37] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14933–14943, 2021. 2

[38] Laura O'Mahony, Vincent Andrearczyk, Henning Müller, and Mara Graziani. Disentangling neuron representations with concept vectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3770–3775, 2023. 3

[39] OpenAI. Hello gpt-4o. https://openai.com/index/hellogpt-4o/, 2024. 4

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[41] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence*, 46(4):2104–2122, 2024. 1

[42] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 2

[43] Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, 2023. 1, 2

[44] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024. 5

[45] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 618–626, 2017. 1, 2

[46] Ao Sun, Pingchuan Ma, Yuanyuan Yuan, and Shuai Wang. Explain any concept: Segment anything meets concept-based explanation. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[47] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021. 7

[48] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations,ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. 2

[49] Teng Wang, Jinrui Zhang, Junjie Fei, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, and Shanshan Zhao. Caption

anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*, 2023. 4

[50] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9677–9696, 2024. 1

[51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2, 4

[52] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 8715–8724, 2020. 1

[53] Wenlong Yu, Ruonan Liu, Dongyue Chen, and Qinghua Hu. Explainability enhanced object detection transformer with feature disentanglement. *IEEE Transactions on Image Processing*, 33:6439–6454, 2024. 7

[54] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021. 2

[55] Shitian Zhao, Zhuowan Li, Yadong Lu, Alan Yuille, and Yan Wang. Causal-cog: A causal-effect look at context generation for boosting multi-modal language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13342–13351, 2024. 2

[56] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2131–2145, 2018. 2, 3, 7