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

Supplementary Material

Contents

S1. Overview of Supplementary Material

In this supplementary material, we mainly give detailed information and analyses about prompt engineering, the CoE approach, and experimental results of the CPE and CoE local explanations. Specifically, as presented in Sec. S2, the prompts that are well-designed in this paper include `prompt-com` for automatically disentangling and describing the commonalities of the given VCs, `prompt-coe` for generating the local explanations given specific samples, and `prompt-coe-eval` for evaluating the local explanations from three explainability metrics. In Sec. S3, we present details of the CoE approach, highlighting its distinguishing qualities. We also discuss the distinctions between CoE and CoT. In Sec. S4, we give experimental results of various versions of CPE, including naive CPE, CPE with clustering, and our final refined version. We also analyze the CPE results explored from different XAI methods (i.e., relevance, activation, and maximum mutual information-based methods) and different model architectures (i.e., ResNet and CLIP). Various examples of VCs, their disentangled concept atoms, probabilities, and CPE values are provided. Besides, we give details on the implementation of the comparison experiments between human evaluations and the CPE metric. In the Section S5, we provide details of the evaluation criteria for linguistic local explanations, along with an overview of the human evaluation process. Details of the comparison between local linguistic explanations generated by different methods and various instances of these local explanations are presented.

Table S1. The prompt template of `prompt-com`. Before querying the LVLMs, we substitute the curly brackets with actual texts.

Prompt for Commonality Disentanglement and Description

```
[System:]
You are a helpful assistant designed to
describe the commonality and specificity
of the given images, and output a JSON
format response.

[User:]
Given {N} images, each containing
highlighted regions, find some common
objects and attributes in these images
and describe each image with words
especially repeated across these images.

Your response should follow these rules:
{Rules}

Your identification process should
follow these steps: {CoT Steps}

Now, please provide your response:
{Response}
```

S2. Prompt Engineering

In this section, we provide a detailed exposition of the three well-engineered prompts designed to describe commonalities of VCs, aggregate all information along the concept circuit to enable the CoE to generate local explanations and evaluate the generated local explanations.

S2.1. Prompt for Commonality Describing

In this paper, we design a sophisticated prompt `prompt-com` to describe the commonalities by a set of concept atoms. The meticulously crafted prompt template is presented in Table S1. As discussed in the main manuscript, this prompt is engineered to accurately

disentangle and summarize the commonalities across multiple subsets of images utilizing precise terminology drawn from 13 semantic directions. Additionally, the disentangled atoms also serve as the foundation for the probability and CPE calculations. To meet these requirements, this prompt incorporates some rules along with step-by-step guidance. The rules outlined below primarily establish 13 semantic directions and delineate the format for output control.

```
  1. Pay more attention to the
repeated objects or attributes across
these images.
  2. Possible objects or attributes
you can use to describe these images
are object category, scene, object
part, color, texture, material,
position, transparency, brightness,
shape, size, edges, and their
relationships.
  3. The identified common objects or
attributes must appear simultaneously
in at least 5 images.
  4. The identified specific objects
or attributes represent some important
contents of an individual image but
not in the common objects or attributes
found in the previous step.
  5. Your description of each image
should be simple and only 3 words.
  6. Your response should be in the
format of a JSON file, of which each
key is a simple image index and each
value is an object or attribute.
```

To enhance the quality of disentanglement and description of atoms, this task is structured into three steps, drawing inspiration from the CoT method.

```
  Step 1, take an overview of all 15
images and summarize all possible
common objects or attributes that
appear simultaneously in at least any
5 of these images.
  Step 2, for each individual image,
identify the common objects or
attributes found in Step 1 that also
appear in the current image to describe
the current image.
  Step 3, for each individual image,
you can also use some specific
attributes or objects that are not
common across these images to describe
the current image if there is not
enough 3-word description for the
common object or attribute found in
Step 2.
```

Table S2. The prompt template of `prompt-coe`. Before querying the LLMs, we substitute the curly brackets with actual texts.

Prompt for CoE Local Explanations

```
[System:]
You are an intelligent deep learning model
explainer and you are now explaining the
decision predicted by a deep vision
identification model.

[User:]
Given a prediction of a deep vision model
and its prediction path formulated in the
format of a concept circuit, you should
first judge whether the model prediction
is correct or incorrect and then give the
reason why the prediction is correct or
incorrect based on the following pieces of
information (A, B, C, D, E). You should
generate an aggregated and rigorous
paragraph based on the given information
rather than imagination.

The information: {A, B, C, D, E}

There are some rules for your response:
{Rules}

Positive and negative prediction examples
are given: {Positive Example}, {Negative
Example}

Your inference process should follow these
steps: {CoT Steps}

Now, please provide your response:
{Response}
```

S2.2. Prompt for CoE Local Explanation

In this paper, we design a prompt `prompt-coe` for the LLM to aggregate all information along the concept circuit and generate a local explanation chain to explain the decision-making process of a DVM. This explanation chain is similar to CoT in terms of the structure of the output. The prompt template is presented in Table S2. The information provided in this prompt includes the DVM's prediction, sample label, image caption, relevant concept explanations derived by applying the CPDF mechanism on the automatically constructed ACD-$\mathcal{B}_{\mathcal{M},\mathcal{T}}$ dataset, and their corresponding relevance values. The concept explanations and relevance values are presented in a structured format.

In this prompt, the rule set primarily functions to regulate the output. Furthermore, we develop two examples of local explanations based on few-shot prompting: a positive example, wherein the CoE generates local explanations corresponding to a correct prediction of the DVM, and a negative example, illustrating the expected local explana-

tion when the DVM prediction is incorrect. Likewise, the process of CoE generating local explanations adheres to the CoT method, as detailed below.

```
Step 1, Based on information A),
which is the model's prediction, and
information B, which is the ground
truth label of the input image, You
first need to determine whether the
two are semantically equivalent.  If
they are semantically equivalent,
then the model's prediction is
considered correct.  If the prediction
and the label are not semantically
equivalent, it is considered an
incorrect prediction.
Step 2, Based on the judgment in
Step 1 and the given information C,
D, and E, which include the caption
of the input image, the vision
model's decision path and the concept
information at each node along the
path, and the concept relevance values
at each node, you need to explain why
the model arrived at this correct
or incorrect prediction.  Analyze
the decision process by examining
each concept in the decision path to
determine how they contributed to the
final outcome.
```

S2.3. Prompt for Evaluating Local Explanations

It is essential to evaluate the generated linguistic local explanations utilizing LVLMs. As illustrated in Table S3, to ensure rigor and precision, we meticulously design a prompt `prompt-coe-eval`, which primarily comprises four components: key information, evaluation criteria, evaluation steps, and rules. The key information includes the image, prediction, label, and the generated local explanations. The three explainability evaluation criteria—Accuracy, Completeness, and User Interpretability—are discussed in detail in Sec. S5.1. Each criterion follows a three-level scoring system (2, 1, 0). These scoring guidelines are explicitly conveyed to the LVLM. The explanation process also adheres to the CoT method, requiring the LVLM to first output the scores alongside evidence and then aggregate these into a final score, as outlined below.

```
Please first provide evidence of
your evaluation for each criterion
and then provide your score for each
criterion, avoiding any potential bias
and ensuring that the order in which
the responses were presented does not
affect your judgment.
```

Table S3. The prompt template of `prompt-coe-eval`. Before querying the LVLMs, curly brackets are filled with actual texts.

Prompt for Evaluating CoE Local Explanations

```
[System:]
You are now a scorer for an
interpretability evaluation system
assessing a deep visual model
interpreter.

[User:]
This interpreter provides natural
language explanations of the
decision-making process of a deep
visual model when given an image
input.
Your task is to evaluate and score
the output explanation of the
interpreter based on specified
criteria to determine its quality.

Your input information includes:
{A, B, C, D}

Based on the four pieces of
information provided above, score
Explanation D according to the
following three criteria. Each
Criteria has its own scoring rules,
and you need to score Explanation D
according to the standards of each
Criteria: {Criteria}

CoT Steps for This Prompt: {CoT
Steps}

Output Control: {Rules}

Now, Please provide your response:
{Response}
```

```
Then sum up the above scores of the
three criteria as the total score.
Finally, output the evidence and
scores for these criteria.
```

S3. Details of the CoE Approach

In this section, we present supplementary details of the CoE approach, including the formulations of the CPE and the distinction from CoT.

S3.1. Probability of Concept Atoms and CPE

In this paper, we acquire the probability of concept atoms by calculating their frequency of occurrence in the disen-

tangled concept atom set $\mathcal{A}$, given a fixed parameter Q and N. The naive version of probability of i_{th} atom is

$$p_i^{Naive} = \frac{Num_i}{Q \times N}. \tag{S1}$$

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

$$H_j^{Naive} = \frac{-\sum_{i=1}^{Q} p_i^{Naive} \log p_i^{Naive}}{\log(Q)}. \tag{S2}$$

As the CPE proposed in this paper serves as an indicator of the interpretability of concepts and DVMs, we normalize the entropy value to a range between 0 and 1 by dividing it by the logarithm of the number total of atoms.

We cluster the atoms in $\mathcal{A}$, as some disentangled atoms are semantically equivalent. The probability and CPE of the clustered atoms in $\mathcal{A}^*$ are formulated as

$$p_i^{Cluster} = \frac{Num_i}{Q \times N}, \tag{S3}$$

$$H_j^{Cluster} = \frac{-\sum_{i=1}^{P^*} p_i^{Cluster} \log p_i^{Cluster}}{\log(P^*)}. \tag{S4}$$

However, there exists a case in which this CPE evaluation becomes ineffective, i.e., when all image patches of a VC are highly similar, as exemplified in the first row of Table S4. The concept should ideally exhibit monosemanticity in this scenario. In contrast, the CPE calculated by Eq. S4 results in a value of 1, as the probabilities of all atoms are evenly distributed (*e.g.*, each with a probability of 1/3 when $P^* = 3$). To mitigate this problem, we set the minimum number of concept atoms in $\mathcal{A}^*$ to N, assuming that each VC contains at least N common atoms. The padding atoms (in a number of $Pad = N – P^*$) are each assigned a frequency of 1. This operation preserves the relative probabilities among the P^* atoms, ensuring that more frequent atoms remain prevalent while less frequent ones retain their lower counts. Upon completion of these procedures, the probability and CPE are updated to Eq. 8 and Eq. 9 in the main manuscript. The experiments are conducted in Sec. S4.1, showing the effectiveness of our method.

S3.2. Explanation of the Concept

In this paper, we define each channel or neuron of DNNs as a VC, represented by a set of masked image patches [7, 10]. Notably, some works consider each image patch as a VC [14], resembling a form of pixel-level semantic segmentation. Channel-based interpretation can serve as both global and local explanations for a DVM by decoding and describing the commonalities among a set of image patches. It better represents the decision concepts learned internally by the DVM. In contrast, the latter, identifying the key regions within a given image, only serves as a local explanation. Our CoE approach can automatically describe these two directions, as both take the form of image patches.

S3.3. Discussion Between CoE and CoT

The CoE approach proposed in this paper draws inspiration from CoT [4, 6], yet with notable distinctions. CoT directly guides large-scale models to articulate their decision-making processes through carefully crafted prompts. However, it has the following limitations: 1. Each step in the CoT still relies on the large model's own capabilities, and each prediction of the current step remains unexplained; 2. CoT mainly emerges in LLMs or LVLMs, whereas the capability for smaller DVMs is insufficient. In contrast, CoE dissects the DVMs by identifying critical decision concepts within key layers. These concepts, described in natural language, serve as nodes in a chain. CoE aggregates these nodes to form a coherent explanation chain that elucidates the DVM's decision-making process. Although the output structure resembles that of CoT, the construction of the CoE explanation chain is achieved by leveraging the general capabilities of LVLMs to automatically describe the VCs and construct the explanation chains. Additionally, CoE provides global conceptual explanations for DVMs while also possessing the capability to quantify polysemanticity. Thus, CoE and CoT are notably distinct.

S3.4. Time and Cost of CoE

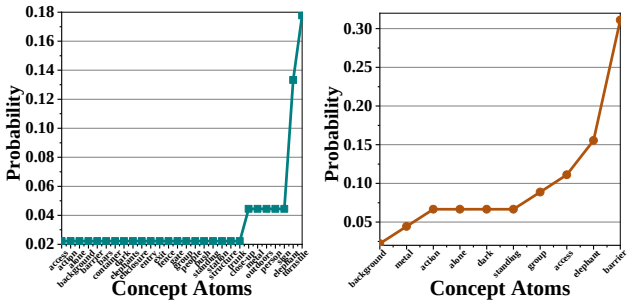
CoE primarily consists of ACD, CPE, CPDF, and local explanation steps. ACD and CPE can be performed offline and obtained through a one-off computation process. Building the global ACD-$\mathcal{B}$ database on ImageNet-val takes 9 hours and costs $70. After that, online inference for the local explanation of a single image requires 20 seconds and costs $0.01. Compared to manual labor, this cost is considered acceptable.

S4. Experiments on CPE

In this section, we provide additional experiments across various versions of CPE, XAI methods, model architectures, and illustrative examples of CPE.

S4.1. Various Versions of CPE

The proposed CPE has evolved through three iterations: the naive version (Eq. S1 and Eq. S2), the clustered version(Eq. S3 and Eq. S4), and the final refined version (Eq. 8 and Eq. 9 in the main manuscript). As shown in Fig. S1, some atoms disentangled for a single concept are semantically equivalent (*e.g.*, barrier and fence, entry and gate), and many of them exhibit low probabilities. After clustering through the entailment model, all semantically redundant atoms are consolidated, leading to adjusted atom probabilities and a reduced CPE value. The semantics of the atoms are mutually exclusive. For a ResNet152 model, 3.5 atoms per concept, on average, are reduced, as illustrated in Fig. S2. Notably, in the third stage, where polysemanticity



(a) Naive Probability (CPE=0.91) (b) Clustered Probability(CPE=0.89)

Figure S1. Examples of two versions of the disentangled concept atom probability distributions. (a) shows the naive version, while (b) represents the clustered one. The channel showed here is number 163 of the output layer of Stage 4 of a ResNet152 model.

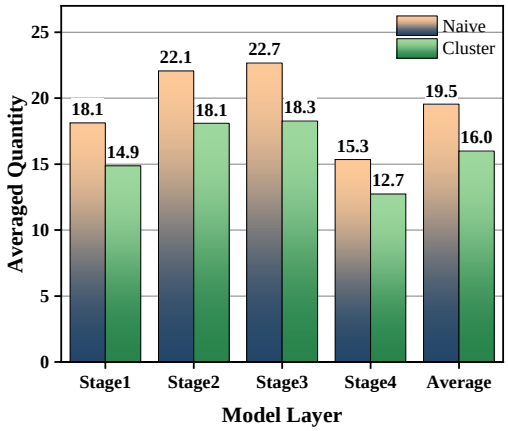


Figure S2. Averaged quantity of disentangled concept atoms. The average term is the averaged quantity of 4 stages.

is most pronounced, the largest reduction is observed, averaging 4.4 per concept. This highlights the significance of introducing the entailment model within the CPDF mechanism to cluster redundant semantics.

Furthermore, as shown in Fig. S2, it is evident that concepts of DVMs exhibit polysemanticity, with the fewest distinct semantics occurring at the final layer (an average of 12.7 non-overlapping semantic atoms) and the most pronounced at stage 3 (18.3). This significantly impairs the interpretability of concepts and DVMs, and the explanations produced by concept-based XAI methods, underscoring the importance of quantifying concept polysemanticity and mitigating its impact on explanations.

As shown in Fig. S3 and exemplified in the first row of Table S4, compared with Fig. 3(a) in the main manuscript, there exist some concepts that require padding. Their commonalities display significant uniformity and an evenly distributed probability pattern, as discussed in Sec. S3.1. This phenomenon, where the actual level of polysemanticity is relatively low but is still calculated as high CPE,

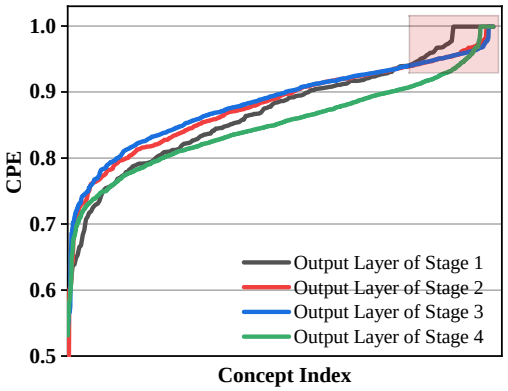


Figure S3. Clustered CPEs along the channel dimension. The channel indices are max-normalized. The magenta-highlighted regions emphasize results where commonalities are relatively limited, yet the computed CPE value is equal to 1.

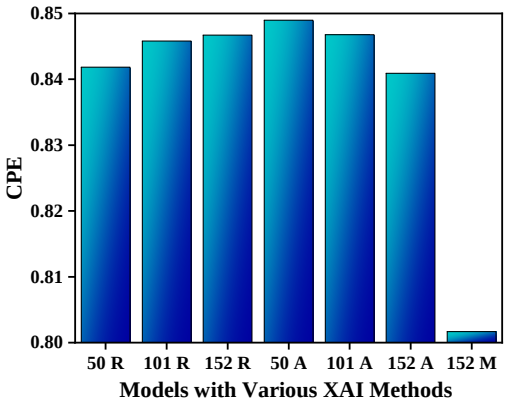


Figure S4. Averaged CPE scores on different XAI methods and different models. The number under the X-axis represents the model depth. R, A, and M stand for Relevance, Activation, and Maximum mutual information-based XAI Methods, respectively.

occurs more frequently in stages 1 and 4 of the DVM. As previously mentioned (discussed in Sec. 4.3 in the main manuscript), these stages are indeed characterized by generally lower levels of polysemanticity in their common concepts. After applying the padding operation, as illustrated in Fig. 3(a) in the main manuscript and the first row of Table S4, the corresponding CPE values drop to relatively low levels. This outcome further substantiates the effectiveness of the CPE proposed in this paper.

S4.2. Different XAI Methods

In this subsection, we conduct experiments on different XAI methods, including the relevance-based [1], activation-based [17], and maximal mutual information-based methods [9]. As illustrated in Fig. S4, the results reveal distinct trends in CPE across different XAI methods. Specifically, the polysemanticity observed in the activation-based

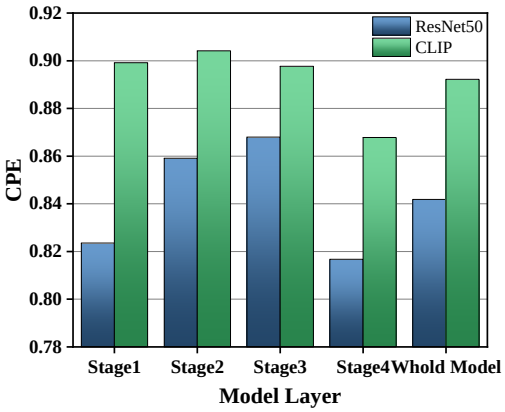


Figure S5. Averaged CPE scores on 2 model architectures, i.e., ResNet50 and CLIP.

method diminishes as model complexity increases, contrasting with the trend exhibited by the relevance-based method. This result indicates that the choice of the XAI method E_x is critical for concept-based explanations. Given that relevance-based concepts achieve superior fidelity and reliability in explanations compared to activation-based concepts [1], this paper adopts the relevance-based CRP as the primary E_x method. Moreover, concepts derived from maximal mutual information exhibit lower CPE values. However, due to the reliance on manually annotated concepts, this method lacks automation and flexibility, limiting its development on DVMs and hindering a more comprehensive comparison with other XAI methods. These results highlight the versatility of our approach in being applicable to various XAI methods and underscore the necessity of automating concept construction.

S4.3. Different Model Architectures

We calculate CPE values across different model architectures, including ResNet50 trained on ImageNet [5, 8] and CLIP-ResNet50 trained on a large-scale vision-language dataset [3, 12, 13]. As presented in Fig. S5, the vision branch of the CLIP model exhibits greater polysemanticity. The general trend aligns with that of the original ResNet, where polysemanticity is lowest in the abstract stage 4 and peaks in the intermediate stages. Polysemanticity is high in the shallowest stage. We infer that these results of CLIP-ResNet50 stem from the constructed global explanation dataset ACD-$\mathcal{B}_{\mathcal{M},\mathcal{T}}$, which is derived from the Out-of-Distributed (OOD) ImageNet Validation dataset $\mathcal{T}$ rather than the independent and identically distributed (iid) vision-language dataset utilized for CLIP's training. Since CLIP operates in a zero-shot mode, the representation of each VC through 15 image patches does not fully align with the conceptual requirements of the original CLIP model, resulting in increased polysemanticity. Moreover, the vision-language dataset utilized for CLIP training encompasses significantly more categories and samples compared to the ImageNet dataset, resulting in an increase in the semantic scope that each concept must represent. This, in turn, amplifies the model's polysemanticity. Our CPE method not only captures and precisely represents these phenomena through the lens of polysemanticity but also proves its effectiveness under zero-shot conditions, reinforcing the validity of the proposed approach.

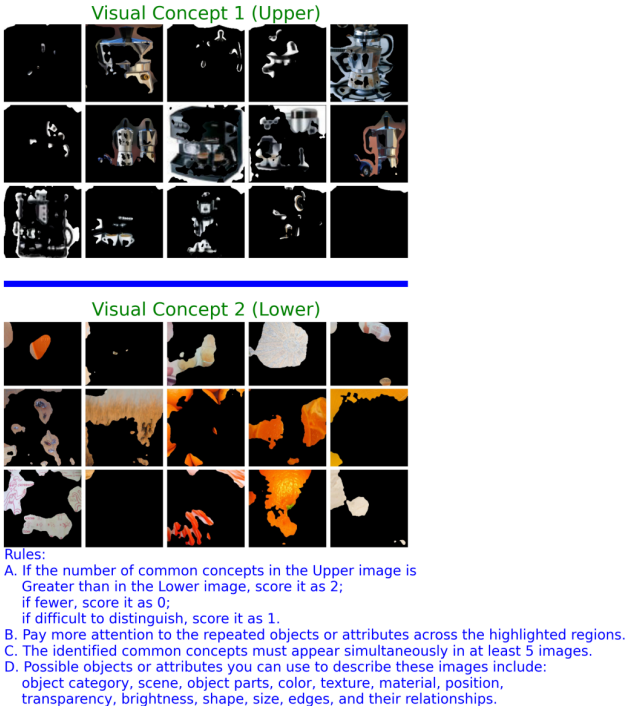


Figure S6. Examples of polysemanticity comparisons between human evaluations and CPE metric.

S4.4. Human Evaluation on CPE

In this subsection, we present a comprehensive overview of the experiments conducted to perform human evaluations on the CPE metric. We first automatically generate a global concept explanation dataset ACD-$\mathcal{B}_{\mathcal{M},\mathcal{T}}$ utilizing the CoE approach proposed in this paper. By incorporating the manually annotated MILAN ANNOTATION dataset [9], we construct a comprehensive VC library for this evaluation, comprising a total of 7,680 VCs. Recognizing the inherent difficulty for humans to directly quantify polysemanticity visually, we instead invite participants to compare the relative degrees of polysemanticity between two VCs. From this VC library, 300 pairs of VCs are randomly sampled, each displaying varying degrees of polysemanticity. As illustrated in Fig. S6, each pair is presented to human evaluators. Participants are instructed to assign a score of 2 if the upper VC exhibits more polysemanticity than the lower

Table S4. Additional examples of VCs and their disentangled concept atoms, along with their concept probability distributions and CPE scores. From top to bottom, the concepts displayed in the table are derived from concepts 141, 366, 1681, and 121 in the fourth stage of a ResNet152 model. The CPE values increase progressively from the top row to the bottom row.

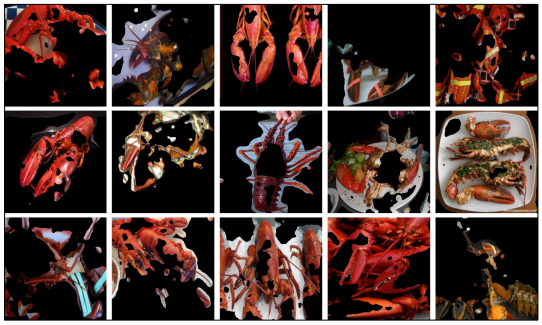
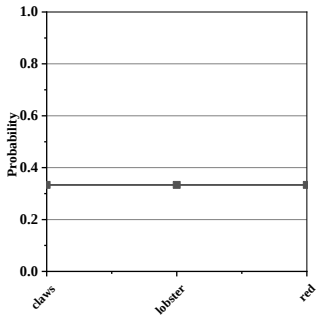
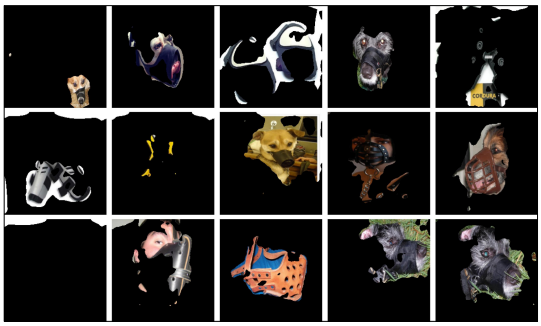
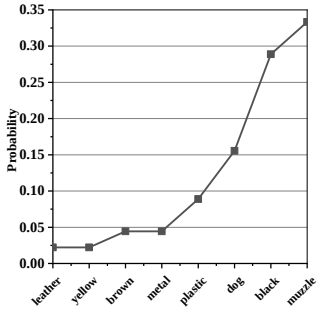
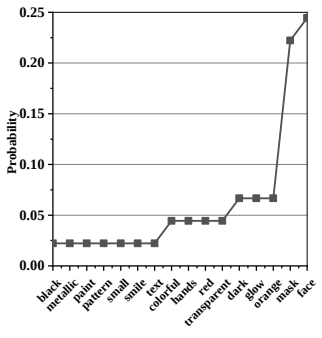
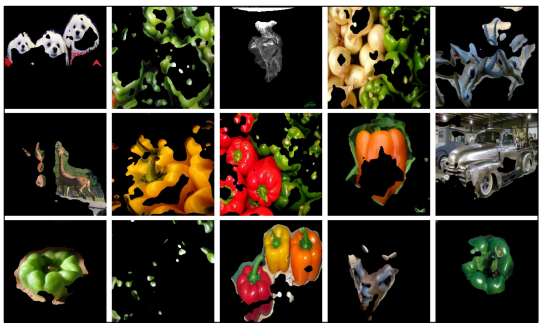
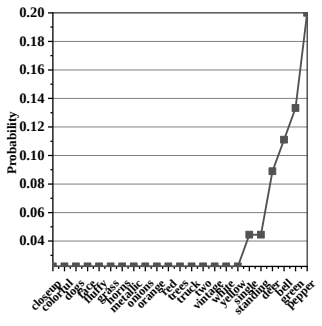
Visual Concept and Disentangled Concept Atoms		Concept Distribution	CPE
	Image 1: [lobster, red, claws], Image 2: [lobster, red, claws], Image 3: [lobster, red, claws], Image 4: [lobster, red, claws], Image 5: [lobster, red, claws], Image 6: [lobster, red, claws], Image 7: [lobster, red, claws], Image 8: [lobster, red, claws], Image 9: [lobster, red, claws], Image 10: [lobster, red, claws], Image 11: [lobster, red, claws], Image 12: [lobster, red, claws], Image 13: [lobster, red, claws], Image 14: [lobster, red, claws], Image 15: [lobster, red, claws]		**0.70**
	Image 1: [dog, muzzle, black], Image 2: [dog, muzzle, black], Image 3: [muzzle, black, plastic], Image 4: [dog, muzzle, black], Image 5: [muzzle, black, plastic], Image 6: [muzzle, black, plastic], Image 7: [muzzle, black, yellow], Image 8: [dog, muzzle, black], Image 9: [muzzle, black, metal], Image 10: [dog, muzzle, brown], Image 11: [muzzle, black, plastic], Image 12: [muzzle, black, metal], Image 13: [muzzle, brown, leather], Image 14: [dog, muzzle, black], Image 15: [dog, muzzle, black]		**0.77**
	Image 1: [orange, glow, dark], Image 2: [face, mask, colorful], Image 3: [face, mask, hands], Image 4: [face, mask, colorful], Image 5: [face, mask, transparent], Image 6: [mask, red, black], Image 7: [mask, metallic, face], Image 8: [orange, glow, face], Image 9: [mask, face, pattern], Image 10: [orange, glow, text], Image 11: [mask, goggles, dark], Image 12: [face, mask, transparent], Image 13: [face, mask, hands], Image 14: [face, paint, smile], Image 15: [dark, small, red]		**0.85**
	Image 1: [white, fluffy, dogs], Image 2: [green, bell, peppers], Image 3: [deer, standing, trees], Image 4: [onions, green, peppers], Image 5: [two, deer, horns], Image 6: [deer, standing, grass], Image 7: [yellow, bell, peppers], Image 8: [red, green, peppers], Image 9: [orange, bell, peppers], Image 10: [metallic, vintage, truck], Image 11: [single, green, pepper], Image 12: [green, bell, peppers], Image 13: [colorful, bell, peppers], Image 14: [deer, face, closeup], Image 15: [single, green, pepper]		**0.88**

Table S5. Explainability metrics utilized for evaluating the local explanations. These criteria are prompted for both GPT-4o and human-based scoring systems. Each criterion has a maximum score of 2 points. The total explanation score is 6. The **bolded** areas represent the core decision rationale for the scoring process.

Metrics	Score	Details of Each Criterion
Accuracy	**2**	**Almost all relevant explanations** focused on key decision points, essential features, important regions, and background information, **with no extraneous or irrelevant content**.
	1	Explanation is **generally relevant but may contain some minor off-topic or unnecessary** information.
	0	Explanation includes a **significant amount of irrelevant** content, diverging from the model's decision-making process and impairing comprehension.
Completeness	**2**	**Comprehensive explanation covering all major steps**, key features, background information, and relevant concepts of the model's decision process.
	1	Explanation addresses **primary decision steps but may slightly overlook some information** or secondary features.
	0	**Incomplete explanation lacking essential decision steps** or information, making comprehension **challenging**.
User Interpretability	**2**	Explanation **allows users without specialized knowledge** to understand the model's decision logic, **with clear, straightforward language and smooth readability**.
	1	Explanation is **mostly understandable to users with a technical background**; it is fairly clear **but may require some re-reading due to less** fluent phrasing or logic.
	0	Explanation is **difficult to comprehend**, with **disorganized or unclear language** that obscures the decision process of the model.

one, a score of 1 if the opposite is true, and 1 if the distinction is unclear. The evaluation guidelines also prompt 13 feasible semantic directions and rules consistent with those prompted for the CPE metric. All VC pairs are divided into 10 groups, with each group evaluated by three participants. If the average score exceeds 1, the upper VC is judged to have greater polysemanticity than the lower one; otherwise, it is assigned a score of 0. Consistency between these results and the CPE metric is then calculated, as summarized in Table 2 in the main manuscript. The results reveal a 75% agreement between human evaluations and the CPE metric, thereby demonstrating the validity of the CPE method.

S4.5. Examples of Disentanglement and CPE

In this subsection, we present additional examples illustrating the disentanglement of VCs into concept atoms, as well as the probability distributions and CPE values of the clustered atoms. As shown in Table S4, the experimental results align with the analyses presented in Sec. 4.2 in the main manuscript. The proposed CoE approach effectively and accurately disentangles VCs into linguistic concept atoms. Furthermore, the entailment model successfully clusters semantically equivalent atoms into mutually orthogonal groups, assigning corresponding probabilities to them. The proposed CPE metric quantifies the polysemanticity of different VCs, with the subjective visual comparisons and the CPE results demonstrating consistent trends. The polysemanticity of the VCs in the table increases progressively from the first row to the last. Correspondingly, the disentangled atoms, their associated probabilities, and the CPE values exhibit consistent changes in alignment with this trend. These results collectively validate the effectiveness of the approach proposed in this paper.

S5. Experiments on CoE Local Explanations

In this section, we elaborate on the evaluation employed to assess the linguistic local explanations. We also present and compare additional instances of local explanations generated by CoE and other methods.

S5.1. Evaluation of Local Explanations

The local explanations are evaluated from three explainability evaluation metrics, namely, Accuracy, Completeness, and User Interpretability [15]. We exclude the fidelity criterion, as CoE finds the key concepts of DVMs through existing concept circuit methods, inherently aligning its fidelity

Table S6. Comparisons of GPT-4o explanation scores under various scenarios. †: the results of baselines are obtained by applying ACD and local explanation steps, without CPDF.

Method	Accuracy	Completeness	User Interpretability	Total Explanation
Places365 Dataset [18] (Baseline†)	1.01	1.07	1.04	3.12
CoE on Places365 Dataset (Ours)	1.68	1.69	1.67	5.04
Chest X-ray Dataset [16] (Baseline†)	1.55	1.62	1.56	4.73
CoE on Chest X-ray Dataset (Ours)	1.81	1.74	1.76	5.31
ViT-B-16 [2] (Baseline†)	1.18	1.14	1.16	3.48
CoE on ViT-B-16 (Ours)	1.65	1.69	1.58	4.92

Table S7. GPT-4o scores on other concept explanation methods.

Method	Acc.	Comp.	User I.	Total
CLIP-Dissect[11] +Descrip.	1.10	1.13	1.08	3.31

with these approaches. Each metric is assigned three score levels: 2 points for optimal performance, 0 points for the lowest performance, and 1 point for a moderate score, as presented in Table S5. The maximum score for each metric is 2, with a total possible score of 6.

These evaluation criteria are provided to both human evaluators and GPT-4o to score the generated linguistic local explanations. To construct the database for GPT-4o-based evaluation, 500 samples are randomly selected from the ImageNet Validation dataset. We sample from the correctly and incorrectly predicted instances of the DVM in a 7:3 ratio, in alignment with the accuracy rate. Three methods, including baseline, CoE without filtering, and CoE, are evaluated in this paper. They generate local explanations for these samples. The evaluation prompt for GPT-4o is discussed in Sec. S2.3. Given the complexity of this evaluation for humans, we randomly select 100 samples from the former database to construct the database for human-based evaluation. The evaluation page, as shown in Fig. S7, consists of the sampled image, the generated local explanations, and the scoring criteria for the three explainability metrics. The three methods are anonymously labeled as Ex1, Ex2, and Ex3. As for human evaluations, the 100 samples are divided into 10 groups, with each group consisting of 30 linguistic explanations assessed by three participants. The results, presented in Table 4 in the main manuscript, demonstrate that the CoE approach outperforms the other two methods across all three explainability metrics, confirming the superiority of the proposed approach.

S5.2. Supplemental Quantitative Evaluation Results

We conduct experiments of CoE on a Transformer architecture (i.e., ViT-B-16) [2]. As shown in the 6th row of Table S6, CoE is effective for the Transformer architecture, achieving an improvement of 1.44 points compared to its

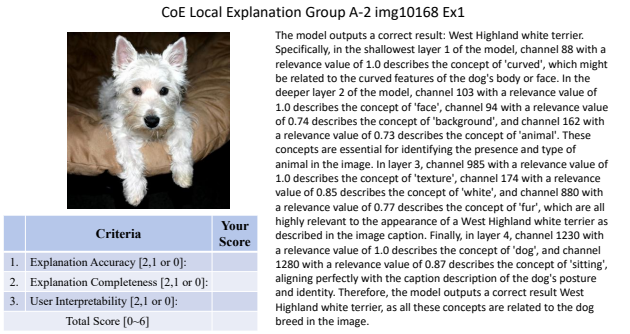


Figure S7. Examples used for evaluating the local explanations generated by three methods.

baseline (the 5th row). Besides, CoE is tested on two other real-world and critical applications (i.e., Places365 [18] and Chest X-ray [16]). Table S6 demonstrates CoE's consistent superiority, achieving scores of 5.04 and 5.31 with improvements of 1.92 and 0.58 over their baselines (without considering polysemanticity). In the medical dataset Chest X-ray [16], CoE achieves an explanation score of 5.31 since the category variety in this dataset is relatively small. All the images depict the human thoracic cavity, and the differences between categories are minimal. This implies that the polysemanticity of concepts learned within the network is more advantageous, enhancing the explainability. We also compare a CLIP-Dissect [11] method for describing concepts, published in ICLR 2023. As shown in Table S7, the overall explainability score is 3.31, which is clearly lower than that of CoE (5.06). CLIP-Dissect generates a single concept atom per channel, which severely underestimates the polysemanticity issue, resulting in insufficient explanations. All results demonstrate the robustness and scalability of the proposed CoE approach.

S5.3. Examples of CoE Local Explanations

In this subsection, we present additional samples to demonstrate the effectiveness of the proposed CoE approach in explaining the decision-making process of DVMs. As presented in Table S8, the experimental results align with the

analyses in Sec. 4.4 in the main manuscript. Nearly all concepts are closely associated with the input images. It infers the logical relationship between the current concept and the entire context according to its contents and relevance values. The final output explanations accurately articulate the decision pathways underlying the DVM's predictions.

S5.4. Comparison Between CoE and Baseline

Additionally, we provide a comparative analysis of CoE-based local explanations with that generated by the baseline method. As exemplified in Table S9, the concepts provided from the baseline method exhibit inconsistencies with the input images, highlighting that disregarding polysemanticity undermines the comprehensibility of local explanations. In contrast, the CoE approach generates linguistic explanations that encapsulate all indispensable relevant information for identifying a hog, such as the concept of pig and pink. These results collectively validate the effectiveness and superiority of the CoE approach proposed in this paper.

Table S8. Additional cases of local explanations generated from the CoE approach. The first three rows show the explanations of correct predictions of the DVM, and the fourth row is reversed. The yellow highlighted regions illustrate the logical relationships identified by the CoE approach between concepts within the explanation chains and between concepts and their contexts. The green highlighted regions represent the CoE approach's final summary and commentary on the explanation chains.

 label: stone wall predict: stone wall	The model outputs a correct result: stone wall. Specifically, in the shallowest layer 1 of the model, channel 88 with a relevance value of 1.0 describes the concept of smooth, and channel 3 with a relevance value of 0.83 describes the concept of ground. These initial detections are fundamental as they set the stage for recognizing more complex features. In the deeper layer 2 of the model, channels such as 401 with a relevance value of 1.0 and 103 with a relevance value of 0.95 both describe the concept of edges, which are crucial for identifying the boundaries and shapes within the stone wall. Channels 24 and 391 with relevance values of 0.91 and 0.84 respectively describe rocks, and channel 26 with a relevance value of 0.73 describes stone, further honing in on the key components of the wall. Channel 506 with a relevance value of 0.72 describes the wall, integrating these features into the overall structure identified in the image. In layer 3, channel 985 with a relevance value of 0.87 describes texture, which is important for recognizing the surface characteristics of the stone wall. Finally, in layer 4, channels 1723 and 1260 with relevance values of 1.0 and 0.81 respectively both describe the wall and stone, confirming the identification of the image as a stone wall. The model's decision path aligns well with the ground truth and the detailed description provided in the caption, leading to a correct prediction.
 label: space shuttle predict: space shuttle	The model outputs a correct result: space shuttle. Specifically, in the shallowest layer 1 of the model, channel 88 with a relevance value of 1.0 describes the concept of 'smooth', which aligns with the smooth surfaces typically seen on space shuttles. In layer 2, channel 103 with a relevance value of 1.0 describes the concept of 'stars', which could relate to the space context of the shuttle, and channel 166 with a relevance value of 0.74 describes 'wooden', which might be less relevant here but could be linked to the indoor display setting of the shuttle. In layer 3, channels 985 and 911 with relevance values of 1.0 and 0.76 respectively describe 'display' and 'background', correctly identifying the shuttle's display context and its background setting in the image. Channels 525 and 233, both describing 'metal' and 'metallic' with relevance values of 0.71 and 0.7, align well with the typical materials used in the construction of a space shuttle. Finally, in the deepest layer 4, channel 1494 with a relevance value of 1.0 accurately identifies the object as a 'shuttle', and channel 625 with a relevance value of 0.85 describes the color 'white', matching the primary color of the shuttle in the image. The model's decision path and the high relevance values of the key concepts directly related to the features of a space shuttle, such as its material and context of display, support the correctness of the prediction. Therefore, the model outputs a correct result: space shuttle.
 label: steel arch bridge predict: steel arch bridge	The model outputs a correct result: steel arch bridge. Specifically, in the shallowest layer 1 of the model, channel 88 with a relevance value of 1.0 describes the concept of orange, which corresponds to the orange sky described in the image caption. In the deeper layer 2 of the model, channel 401 with a relevance value of 1.0 describes the concept of curved, channel 404 with a relevance value of 0.75 describes the concept of sunset, channel 182 with a relevance value of 0.72 describes the concept of water, and channel 103 with a relevance value of 0.71 describes the concept of yellow. These concepts align well with the image description, capturing the curved shape of the bridge, the sunset, and the water reflection. In layer 3, channel 395 with a relevance value of 1.0 describes the concept of cloud, and channel 725 with a relevance value of 0.76 further supports the concept of sunset, enhancing the contextual understanding of the scene. Finally, in layer 4, channels 1120 and 573 both with high relevance values (1.0 and 0.92 respectively) describe the concept of bridge, directly identifying the main object in the image. The structured and high relevance of these concepts at each layer demonstrates a clear and accurate recognition path from general scene attributes to the specific identification of the steel arch bridge, confirming the correctness of the model's prediction.
 label: gar predict: pencil box	The model outputs an incorrect result pencil box instead of the correct label gar. Specifically, in the shallowest layer 1 of the model, channel 88 with a relevance value of 1.0 describes the concept of orange, which may relate to the orange lines mentioned in the image caption. In the deeper layer 2, channel 229 with a relevance value of 1.0 correctly identifies the concept of fish, aligning with the caption that describes a large fish. Moving to layer 3, channels 422 and 31 with relevance values of 1.0 and 0.99 respectively continue to support the identification of the object as a fish and an animal. However, channel 519 with a relevance value of 0.82 describes the concept of measurement, likely influenced by the presence of a ruler in the image, which is a common object in a pencil box. In layer 4, channel 1147 with a relevance value of 1.0 describes the concept of black, possibly picking up on the black cap mentioned in the caption, and channel 626 with a relevance value of 0.74 describes the concept of ruler. The model seems to have been misled by the presence of objects typically found in a pencil box (ruler, black cap) and the concept of measurement, leading it to incorrectly predict the image as a pencil box rather than recognizing the primary subject as a gar fish. Therefore, the model outputs an incorrect result pencil box.

Table S9. Comparisons between local explanations generated by the baseline method and our CoE approach. The highlighted regions indicate the key concepts identified by different methods as influential in the model's decision-making process. The local explanations in the first row are generated by the baseline method using the manually annotated MILAN ANNOTATION dataset, while the second row represents the outputs of our CoE approach applied to the automatically constructed ACD-$\mathcal{B}_{\mathcal{M},\mathcal{T}}$ dataset.

MILAN ANNO. (Baseline)	 label: hog predict: hog	The model outputs a correct result hog. Specifically, in the shallowest layer 1 of the model, channel 88 with a relevance value of 1.0 describes the concept of a bear running across sand and a dog going through the pasture with flowers, which might not directly relate to the image of piglets but could be triggered due to the presence of animals in motion or in a pastoral setting. In the deeper layer 2 of the model, channel 46 with a relevance value of 0.82 describes the concept of animals, which directly correlates to the image content of piglets. Furthermore, in layer 4 of the model, channel 475 with a relevance value of 0.75 describes the concept of pictures of animals, which strongly supports the correct identification of the image as containing animals, specifically hog. The decision path and relevance values indicate that the model has effectively recognized the primary subject (animals) in the image, leading to the correct prediction of hog, aligning with the ground truth label and the image caption describing piglets.
CoE (Ours)	 label: hog predict: hog	The model outputs a correct result: hog. Specifically, in the shallowest layer 1 of the model, channel 88 with a relevance value of 1.0 describes the concept of 'smooth', which may relate to the texture of the piglets in the image. In the deeper layer 2 of the model, channels 391, 506, 46, 26, and 24 with relevance values of 1.0, 0.88, 0.82, 0.77, and 0.75 respectively, all describe the concept of 'animal', aligning well with the presence of piglets in the image. Channel 229 with a relevance value of 0.77 describes the concept of 'grid', which corresponds to the wire fence enclosure mentioned in the caption. Channel 402 with a relevance value of 0.71 describes 'food', which is consistent with the description of one piglet eating. In layer 3, channel 985 with a relevance value of 1.0 continues to describe 'grid', reinforcing the identification of the enclosure. Channels 591 and 498 with relevance values of 0.96 and 0.78 describe 'creature' and 'animal' respectively, further confirming the presence of living animals in the image. Finally, in layer 4, channels 726 and 475 with relevance values of 1.0 and 0.75 describe 'pig', and channel 1884 with a relevance value of 0.84 describes 'pink', directly correlating with the pink piglets observed. These concepts are all highly relevant to the image's content as described in the caption, leading to a correct prediction of 'hog'.

References

[1] 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. 5, 6

[2] Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Attnlrp: attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning*, pages 135–168, 2024. 9

[3] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 6

[4] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. A survey of chain of thought reasoning: Advances. *Frontiers and Future*, 2023. 4

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 6

[6] 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. 4

[7] Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2024. 4

[8] 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. 6

[9] 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. 5, 6

[10] 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. 4

[11] Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. *arXiv preprint arXiv:2204.10965*, 2022. 9

[12] 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. 6

[13] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 6

[14] 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. 4

[15] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021. 8

[16] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 9

[17] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 5

[18] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 9