WISE: Weak-Supervision-Guided Step-by-Step Explanations for Multimodal LLMs in Image Classification

Yiwen Jiang [1,2] **Deval Mehta** [1,2,†] **Siyuan Yan** [1,2] **Yaling Shen** [2]
Zimu Wang [2] **Zongyuan Ge** [2,†]
[1]Faculty of Engineering, Monash University, Melbourne, Australia
[2]AIM for Health Lab, Faculty of IT, Monash University, Melbourne, Australia
{yiwen.jiang, deval.mehta, zongyuan.ge}@monash.edu

Abstract

Multimodal Large Language Models (MLLMs) have shown promise in visual-textual reasoning, with Multimodal Chain-of-Thought (MCoT) prompting significantly enhancing interpretability. However, existing MCoT methods rely on rationale-rich datasets and largely focus on inter-object reasoning, overlooking the intra-object understanding crucial for image classification. To address this gap, we propose WISE, a Weak-supervIsion-guided Step-by-step Explanation method that augments any image classification dataset with MCoTs by reformulating the concept-based representations from Concept Bottleneck Models (CBMs) into concise, interpretable reasoning chains under weak supervision. Experiments across ten datasets show that our generated MCoTs not only improve interpretability by 37% but also lead to gains in classification accuracy when used to fine-tune MLLMs[1]. Our work bridges concept-based interpretability and generative MCoT reasoning, providing a generalizable framework for enhancing MLLMs in fine-grained visual understanding.

1 Introduction

Deep Learning (DL) models have achieved remarkable performance, powering applications in various domains. However, DL architectures are inherently "black-box" which often result in limited interpretability of the underlying decision-making processes (Papernot et al., 2017). Thus, an increasing amount of attention has been directed toward developing DL models that are either inherently interpretable, or capable of producing explicit chains of reasoning that reveal their conclusions.

In the realm of generative DL, very recent Multimodal Large Language Models (MLLMs) have become a powerful paradigm for joint visual-textual processing and importantly introducing reasoning (Liu et al., 2023) to decision-making. To enhance their interpretability, recent work has proposed Multimodal Chain-of-Thought (MCoT) reasoning (Zhang et al., 2024b), which simulates human-like step-by-step inference by breaking down complex problems into sequential, interpretable reasoning steps. This has proved to improve interpretability via reasoning, which ultimately enhances the performance on multimodal tasks (Chen et al., 2024).

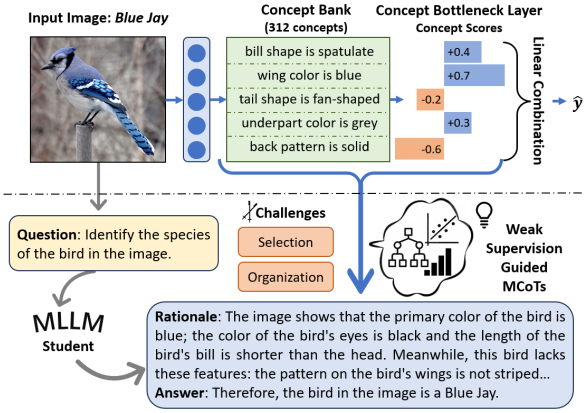

Figure 1: Exhaustive concept sets and the intrinsic linear combination in CBMs hinder their direct transformation into CoT. Our method addresses these challenges through a weak-supervision-guided reformulation of the bottleneck layer into concise textual rationales.

Datasets with rationales are essential for eliciting MCoT reasoning in MLLMs. However, they are typically constructed through costly human annotations or automatic generation via prompting LLMs, both of which pose challenges in ensuring high data quality. Existing MCoT datasets, such as CoMT (Cheng et al., 2025), primarily target complex inter-object reasoning, and none to date focus on image classification tasks involving intra-object understanding. However, prior work (Zhang et al., 2024a) has shown that integrating classification-focused data into MLLMs training, even without MCoT supervision, can enhance higher-level visual

†Corresponding authors.
[1]Data and codes are available on: https://github.com/yiwenJG/WISE-MCoT

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 14674–14685
November 4-9, 2025

capabilities such as visual question answering and reasoning, highlighting classification as a foundation for multimodal reasoning. Since many MCoT approaches (Chen et al., 2023) begin by detecting objects in the image, classification naturally contributes to reasoning accuracy. However, the problem of generating high-quality natural language MCoT tailored to image classification through automated methods remains largely unexplored.

Parallel to recent efforts in generative DL, research on discriminative models has predominantly relied on concept-based analysis (Kim et al., 2018), exemplified by Concept Bottleneck Models (CBMs) (Koh et al., 2020), to enhance interpretability (Mehta et al., 2025). CBMs aim to associate each neuron with a human-understandable concept. In image classification, they map visual representations to a set of textual concepts, from which predictions are derived via a linear combination of concept scores. CBMs also offer intervenability: Within a human-in-the-loop framework (Yan et al., 2023b), humans can alter predictions by adjusting wrongly activated concepts, enabling direct control over model behavior (Koh et al., 2020). Recent CBM work has pushed the paradigm forward by enabling fully automated language grounding. (Oikarinen et al., 2023; Yang et al., 2023; Yan et al., 2023a), which prompts pre-trained Large Language Models (LLMs) with category names to generate candidate concepts, selects representative ones to construct a concept bank, and employs multimodal models such as CLIP (Radford et al., 2021) to align images and concepts via image-text scoring, forming a Concept Bottleneck Layer (CBL).

Inspired by the success of CBMs in interpretable image classification, we pose a natural question: **Can the CBL be reformulated as MCoT to facilitate the training of MLLMs?** A naive approach that directly transforms concepts into natural language is not feasible due to the extensive set of possible concepts and the importance of their ordering (Figure 1). Addressing this question requires overcoming two key challenges: (1) Selecting appropriate concepts to serve as components of the MCoT. Unlike CBMs, which score and combine all concepts during inference, generative models cannot feasibly incorporate such exhaustive representations. For instance, the CUB dataset (Wah et al., 2011) contains 312 annotated bird attributes, making it impractical to reflect all CBM neurons in a single rationale; (2) Organizing the selected concepts into coherent reasoning chains that align with human cognitive patterns. In CBMs, each concept may contribute positively or negatively to a prediction, supporting or refuting specific categories. Moreover, concepts vary in their contribution to the final decision. These aspects must be carefully modeled to construct effective MCoTs.

Motivated by the principle of weak-to-strong generalization (Burns et al., 2024), we propose a novel Weak-supervIsion-guided Step-by-step Explanation method (**WISE**) for automatic MCoT generation. It reformulates the CBL as concept-driven natural language reasoning. Specifically, we use CLIP to score images against a concept bank and utilize CBMs for concept annotation. Leveraging the prior distribution between categories and concepts, we apply decision tree (Breiman et al., 1984) to construct Prior Trees. To capture instance-level variation and reflect the human tendency to organize concepts sequentially, we further design two instance-specific trees: an Affirmation Tree and an Elimination Tree. These trees are then combined and transformed into MCoTs. Finally, we fine-tune MLLMs using a curriculum learning strategy. Experiments across ten image classification datasets show that our method improves the interpretability of MLLMs by 37% and enhances classification accuracy, despite being guided by models that are more lightweight than billion-parameter MLLMs.

Furthermore, our generated MCoTs align with human reasoning patterns and directly address the two previously identified challenges: by incorporating both discriminativeness and visual salience, they focus on a small set of critical concepts for decision-making; they also capture concept typicality and reflect both affirmative and counterfactual reasoning strategies. Overall, our main contributions are as follows:

- To the best of our knowledge, we are the first to bridge the previously separate paradigms of CBMs and MCoTs by proposing WISE, a weak-supervision-guided method that reformulates CBM representations into natural language MCoTs.
- Our method transforms any image classification dataset with category labels into an MCoT-augmented version, producing rationales that reflect human reasoning by integrating category typicality, instance-level distinctiveness, and both supportive and counterfactual evidence.
- We conduct experiments on ten image classification datasets, showing that our generated MCoTs improve the interpretability of MLLMs by 37% while also enhancing classification accuracy.

2 Related Work

Concept Bottleneck Models. CBMs (Koh et al., 2020) are a prominent approach for designing inherently interpretable DL models, as detailed by Zhou et al. (2018) and Losch et al. (2019). CBMs incorporate a concept bottleneck layer preceding the final fully connected layer, where each neuron represents a human-interpretable concept. Yüksekgönül et al. (2023) and Oikarinen et al. (2023) proposed data-efficient methods to convert any DL models into CBMs without training from scratch. CLIP-based CBMs (Jiang et al., 2025; Shang et al., 2024; Oikarinen et al., 2023; Yang et al., 2023; Yan et al., 2023a) have leveraged vision-language alignment learned during pre-training (Radford et al., 2021) to eliminate the need for concept-level manual annotation, enabling automatic concept bank construction via LLMs and achieving competitive performance with "black-box" models. Recent studies such as Concept Agent (Jiang et al., 2025) and LM4CV (Yan et al., 2023a) explores concise concept banks to reduce redundancy, but the intrinsic CBM architecture, the linear combination of individual concept scores, limits their direct use in CoT reasoning for MLLMs. Crucially, our method dynamically selects concepts for reasoning on a per-image basis, instead of relying on a predefined, fixed set.

MCoT Reasoning for MLLM. Multimodal Chain-of-Thought (MCoT) extends the Chain-of-Thought (CoT) paradigm (Wei et al., 2022) to MLLMs, aiming to enhance their ability to perform stepwise reasoning across diverse input modalities. CoT improves both transparency and accuracy by decomposing complex problems into intermediate steps (Zhang et al., 2024b). In explainable image classification, MCoT incorporates visual inputs while generating rationales in natural language. To better structure the reasoning process, various topologies such as trees (Yao et al., 2023) and graphs (Besta et al., 2024) have been explored, enabling richer semantic composition and flexible backtracking. Prompt-based methods enable MLLMs to produce rationales at inference time via crafted instructions or in-context examples, requiring no additional training (Luo et al., 2025; Zheng et al., 2024; Gao et al., 2024). In contrast, learning-based methods (Zhang et al., 2024b) fine-tune MLLMs on annotated rationale data, making them more effective at implicitly acquiring reasoning patterns. Although several MCoT datasets are available (Chen et al., 2023), they primarily target reasoning over inter-object relations or rudimentary knowledge. To date, there remains a lack of rationale-annotated datasets tailored to image classification that emphasize intrinsic properties of individual objects.

Weak-to-Strong Interpretability. Weak-to-strong generalization (Burns et al., 2024) is a paradigm for eliciting the latent capabilities of powerful models through supervision provided by weaker models. Building on this idea, we explore whether weak supervision, including CBM (Koh et al., 2020), CLIP (Radford et al., 2021), Decision Tree (Breiman et al., 1984), Linear Regression (Hastie et al., 2009), and Bayesian Learning (Bayes, 1958), can elicit interpretable MCoT reasoning from MLLMs. Among these, decision trees play a central role, recursively splitting the feature space based on input attributes to form a tree-like structure (Costa and Pedreira, 2023) where internal nodes represent decision rules and leaf nodes correspond to outcomes. This transparent structure enables intuitive interpretation of the decision-making process and has been extensively studied for applications in critical domains such as healthcare (Jiang et al., 2023).

3 Methodology

3.1 Problem Formulation

Given a dataset $\mathcal{D} = (x_i, y_i)_{i=1}^{\mathcal{K}}$, where $x_i \in \mathcal{X}$ denotes an image and $y_i \in \mathcal{Y}$ is one of $\mathcal{N}$ predefined categories, and a concept set $\mathcal{C} = c_1, c_2, \ldots, c_M$ provided by humans or LLMs as the semantic basis for interpretation, the objective is to automatically generate textual rationales R to guide the fine-tuning of MLLMs on $\mathcal{D}$.

We formulate this as modeling the joint distribution $P(\mathcal{Y}, R \mid \mathcal{X}, Q)$, where Q denotes a task-specific prompt (e.g., *Identify the species of the bird in the image*). The rationales R are MCoTs constrained to reflect a subset of the concept set, i.e., $\text{Info}(R) \subseteq \mathcal{C}$, with $|\text{Info}(R)| \ll M$, enabling concise and faithful concept-driven reasoning. $\mathcal{Y}$, R and Q are represented as a sequence of language tokens in MLLMs.

3.2 Concept Scoring for Visual Salience

Humans tend to prioritize visually salient features as key cues for inference when recognizing objects. Motivated by this observation, we initiate rationale generation using Visual-Language Models (VLMs), such as CLIP (Radford et al., 2021). CLIP consists of two encoders: an image encoder $\mathcal{I} : \mathcal{X} \rightarrow \mathcal{R}^d$ and a text encoder $\mathcal{T} : \mathcal{C} \rightarrow \mathcal{R}^d$,

which project images and textual concepts into a shared d-dimensional embedding space. The concept score between an input image x_i and a concept c_m is calculated using cosine similarity:

$$s(x_i, c_m) = \mathcal{I}(x_i) \cdot \mathcal{T}(c_m) \tag{1}$$

This score quantifies the degree of cross-modal alignment and reflects the visual salience of concept c_m in image x_i. To obtain concept-class alignment weights, we treat the concept score vector $s(x_i) \in R^M$ as a concept bottleneck layer and apply a softmax classifier trained with label supervision. This allows us to learn a weight vector $\mathbf{w}_{y_n} \in R^M$ for each class y_n, where each dimension reflects the relative importance and polarity (positive or negative) of a concept for that class:

$$P(y_n \mid x_i) = \frac{\exp(\mathbf{w}_{y_n}^\top s(x_i) + b_n)}{\sum_{j=1}^{\mathcal{N}} \exp(\mathbf{w}_{y_j}^\top s(x_i) + b_j)} \tag{2}$$

Here, $\mathbf{w}_{y_n}$ and b_n denote the weight vector and bias term for class y_n. Finally, for an image x_i labeled as y_n, we derive the binary annotation for each concept c_m based on the sign of its contribution to the predicted class. Let $z_{i,m} \in \{0, 1\}$ denote the annotation of concept c_m in x_i:

$$z_{i,m} = \begin{cases} 1, & \text{if } s(x_i, c_m) \cdot \mathbf{w}_{y_n}^{c_m} > 0 \\ 0, & \text{if } s(x_i, c_m) \cdot \mathbf{w}_{y_n}^{c_m} \leq 0 \end{cases} \tag{3}$$

For datasets which possess the ground-truth concept labels, the binary annotation step is not required. We then apply logistic regression, supervised by these labels, to map concept scores to probabilities, denoted as $P_{i,m}$ and determine a per-concept threshold optimized for macro F1 score. Concept instances with probabilities below the threshold, even if they are true positives, are relabeled as negative. This refinement retains only the most visually salient concepts.

3.3 Category Typicality Tree Modeling

During object recognition, humans often rely on stereotypical impressions associated with imagined categories. From a Bayesian perspective, this corresponds to a prior probability (Bayes, 1958), specifically the category-to-concept prior denoted as $P(c_m \mid y_n)$ in our setting. To model this prior knowledge, we compute the prior distribution for each category–concept pair as follows:

$$P(c_m \mid y_n) = \frac{1}{|\mathcal{D}_{y_n}|} \sum_{i \in \mathcal{D}_{y_n}} P_{i,m} \tag{4}$$

Here, $\mathcal{D}_{y_n} = \{i \mid y_i = y_n\}$ denotes the set of training instances labeled with class y_n. This computation yields a prior matrix $\mathbf{P} \in \mathcal{R}^{N \times M}$, where each entry $[\mathbf{P}]_{n,m}$ represents the prior strength linking class y_n and concept c_m. We use this matrix to build a category-specific decision tree for each y_n.

Prior Tree. The objective of the tree modeling is to discover the shortest decision path that distinguishes y_n from the remaining classes, which are treated as negative samples. To this end, we identify the most salient concepts for y_n by filtering the concept dimensions with $[\mathbf{P}]_{n,m} > 0.5$. These selected concepts serve as input features for the decision tree algorithm, which recursively select the concept that yields the highest information gain according to Gini impurity (Breiman et al., 1984) at each node. The resulting decision path for class y_n is denoted as:

$$\mathcal{T}_p(y_n) = \{c_1, c_2, \ldots, c_p\} \tag{5}$$

where $\mathcal{T}_p(y_n)$ is an ordered sequence of concepts forming the decision path, and p denotes its length.

3.4 Instance Distinctiveness Tree Modeling

While stereotypical impressions reflect the prototypical characteristics of a category, individual instances often exhibit distinctive features, i.e., deviations from the prototype. Such variations are essential for human reasoning. We capture this distinctiveness by also formulating a tree-based learning framework. We posit that human reasoning often follows a two-step process to account for variations and organization of multiple concepts: (1) it begins with affirmation, supporting a hypothesis based on observed concepts, and (2) when evidence is insufficient, it proceeds to elimination, ruling out confounding options based on absent concepts. To reflect this strategy, we decompose our tree construction into two sequential stages.

Affirmation Tree. This tree builds upon the Prior Tree and focuses exclusively on positive concepts supporting the target class. For a given instance (x_i, y_n), we begin by identifying the subset of prior concepts that are both part of the category-level decision path $\mathcal{T}_p(y_n)$ and observed in x_i. We denote this instance-specific subpath as:

$$\mathcal{T}_p^+(x_i, y_n) = \{c \in \mathcal{T}_p(y_n) \mid z_{i,c} = 1\} \tag{6}$$

We then retrieve a set of hard negative instances from $\mathcal{D}$—instances not labeled as y_n but sharing the same set of activated prior concepts $\mathcal{T}_p^+(x_i, y_n)$.

Since these instances cannot be distinguished from x_i using the Prior Tree alone, additional instance-specific concepts are needed to further support.

To resolve this ambiguity, we extract additional concepts from x_i that are not included in the prior decision path $\mathcal{T}_p(y_n)$ but are present in the instance. These distinctive, instance-specific concepts serve as input features for constructing the Affirmation Tree, and their resulting decision path is denoted as:

$$\mathcal{T}^{+}_{\text{ins}}(x_i, y_n) = \{c_1, c_2, \ldots, c_{\text{ins}^+}\} \tag{7}$$

where $c_j \in \mathcal{C} \setminus \mathcal{T}_p(y_n)$ and $z_{i,c_j} = 1$. Finally, the Affirmation Tree path is defined as:

$$\mathcal{T}^{+}_{a}(x_i, y_n) = \mathcal{T}^{+}_{p}(x_i, y_n) \parallel \mathcal{T}^{+}_{\text{ins}}(x_i, y_n) \tag{8}$$

Here, $\parallel$ denotes path concatenation. The resulting concepts within the path is reordered according to their prior probabilities that reflect relative importance. In cases where the leaf node of the Affirmation Tree $\mathcal{T}^{+}_{a}(x_i, y_n)$ retain non-zero Gini impurity, an Elimination Tree is constructed to further disambiguate confounding classes.

Elimination Tree. Let y_c denote the set of confounding classes that causes non-zero Gini impurity at the leaf node of the Affirmation Tree for the input x_i. The goal of the Elimination Tree is to exclude these classes by leveraging concepts that are absent in x_i. Specifically, we collect negative instances as $\mathcal{D}_{y_c} = \{i \mid y_i \in y_c\}$ and identify true negative concepts in x_i as input features for decision tree construction. The resulting decision path is defined as:

$$\mathcal{T}^{-}_{e}(x_i, y_n) = \{c_1, c_2, \ldots, c_{\text{ins}^-}\} \tag{9}$$

where $c_j \in \mathcal{C}$ and $z_{i,c_j} = 0$. These concepts are absent in x_i but frequently occur in the confounding classes y_c, thereby complementing the insufficient evidence provided by the Affirmation Tree. If non-zero Gini impurity persists after applying both trees, we recommend expanding the concept bank $\mathcal{C}$, indicating that the current set of concepts is insufficient to distinguish between categories.

3.5 Tree-Guided Rationale for MLLMs

We define the final MCoT decision path as the concatenation of the Affirmation Tree and the Elimination Tree for a given instance (x_i, y_n):

$$\mathcal{T}_{\text{MCoT}}(x_i, y_n) = \mathcal{T}^{+}_{a}(x_i, y_n) \parallel \mathcal{T}^{-}_{e}(x_i, y_n) \tag{10}$$

The resulting chain captures both supportive and exclusionary reasoning steps, accounting for both category-level prototypicality and instance-level distinctiveness. To convert this structured path into natural language rationales, we design a template-based generation module that verbalizes each concept c_k into a descriptive clause, conditioned on its semantics and polarity. As shown in Figure 2, the resulting clauses are then sequentially composed into a coherent explanation that mirrors the underlying reasoning logic. Future work may explore leveraging generative LLMs to polish and rephrase the rationales. In our experiments, we retain the template-based method to facilitate concept clause extraction and evaluation via regular expressions.

3.6 Fine-tuning MLLMs with MCoTs

Rather than immediately fine-tuning the MLLMs to perform concept-driven reasoning, we begin with task adaptation to guide the model in learning how to ground individual concepts in images. Following the principle of curriculum learning (Bengio et al., 2009), we adopt a two-stage fine-tuning strategy that gradually increases task complexity.

In the first stage, we create a question-answering dataset by templating each annotated concept into a natural language QA pair (e.g., *Q*: *What color are the bird's feathers? A*: *Blue*), enabling the model to associate visual features with individual concepts. In the second stage, we further fine-tune the model on the Tree-guided MCoT dataset to enable compositional reasoning over multiple concepts.

4 Experiments

4.1 Datasets

We evaluate our method on ten fine-grained image classification datasets across various domains and scales, including CUB (Wah et al., 2011), SkinCon (Daneshjou et al., 2022), LAD (Zhao et al., 2019), Oxford-Flowers (Nilsback and Zisserman, 2008), and Oxford-Pets (Parkhi et al., 2012). Among these, CUB, SkinCon, and LAD offer image-level concept annotations that enable quantitative evaluation of MCoT interpretability.

CUB contains 200 bird species, with each image annotated using 312 binary concept labels. SkinCon is a dermatology dataset with hierarchical skin disease labels. Following Daneshjou et al. (2022) and Pang et al. (2024), we construct two variants: SkinCon (3-class), using three coarse-grained disease categories and 22 concepts that appear in at least 50 images; and SkinCon (9-class), covering nine fine-grained categories with all 48 concepts.

Model		Weak Supervisors				MLLM (Qwen2-VL)			
Dataset		CLIP	CBM	DT	NBC	ZS-IO	ZS-MCoT	IT-IO	**IT-MCoT (ours)**
CUB	acc.	59.30	45.94	28.65	60.30	33.21	26.89	82.40	83.69 (+1.29)
	intp.	-	55.62	87.79	50.51	-	55.20	-	65.03
SkinCon-3	acc.	64.71	68.42	56.97	68.42	67.18	57.89	74.61	74.61 (-)
	intp.	-	50.82	79.92	56.64	-	24.33	-	87.40
SkinCon-9	acc.	11.76	60.06	60.06	60.06	22.60	25.70	60.37	62.23 (+1.86)
	intp.	-	50.15	91.67	53.55	-	28.76	-	93.45
LAD-A	acc.	96.83	73.36	57.41	86.63	79.67	82.49	97.50	97.62 (+0.12)
	intp.	-	60.18	82.89	50.45	-	55.90	-	82.78
LAD-E	acc.	75.56	51.94	44.08	76.84	63.30	67.53	94.55	94.70 (+0.15)
	intp.	-	55.92	82.12	51.49	-	69.44	-	98.74
LAD-H	acc.	32.49	13.88	9.75	19.04	21.17	21.20	59.02	61.61 (+2.59)
	intp.	-	57.50	60.38	51.52	-	23.28	-	83.44
LAD-F	acc.	73.33	43.71	36.06	61.64	70.14	70.05	93.47	93.17 (-0.30)
	intp.	-	56.23	76.77	50.65	-	74.41	-	96.67
LAD-V	acc.	71.68	51.21	42.34	75.25	47.92	48.30	94.10	93.88 (-0.22)
	intp.	-	58.57	75.61	52.04	-	74.70	-	95.12
Average	acc.	60.71	51.07	41.92	63.52	50.65	50.23	82.00	**82.69** (+0.69)
	intp.	-	55.62	79.64	52.11	-	50.75	-	**87.83**

Table 1: **Main results** on 8 concept-annotated datasets, reporting *classification accuracy* (acc.) and *interpretability* (intp.) for weak supervisors and MLLM. *CLIP* is used in a zero-shot setting. The remaining methods are CLIP-based weak supervisors: Concept Bottleneck Model (*CBM*), Decision Tree (*DT*), and Naive Bayes Classifier (*NBC*). *ZS-IO* and *ZS-MCoT* denote zero-shot input-output QA and zero-shot MCoT. *IT-IO* is instruction tuning without rationales; *IT-MCoT (ours)* integrates MCoTs derived from weak supervision. Accuracy gains over the *IT-IO* baseline are indicated in green (improvement) and red (decline). Top-1 and Top-2 interpretability scores are highlighted in blue.

LAD consists of five sub-datasets covering animals (LAD-A), electronics (LAD-E), hairstyles (LAD-H), fruits (LAD-F), and vehicles (LAD-V). It includes 230 categories and 359 concepts in total. Following Jiang et al. (2025), we construct concept banks for Oxford-Flowers and Oxford-Pets by prompting GPT-4o (OpenAI et al., 2024).

4.2 Experimental Details

Weak Supervisors. We evaluate both accuracy and interpretability of four weak supervisors, which are used to guide MLLMs in generating MCoTs: (1) **CLIP-zero-shot** (Radford et al., 2021) Serves as a baseline to showcase the pretrained vision-language alignment and classification capabilities of CLIP. (2) **CLIP-based CBM** (Yüksekgönül et al., 2023, Oikarinen et al., 2023, Yan et al., 2023a) Applies logistic regression on CLIP-derived concept scores. Interpretability is quantified by the polarity of the product between concept weights and scores, indicating each concept's contribution on the prediction. (3) **CLIP-based Decision Tree** (Breiman et al., 1984) Constructs decision rules over CLIP-annotated concepts. As the tree lacks direct visual access, interpretability is measured by the correctness of CLIP concept labels along decision paths. When annotations are fully accurate, the explanation is exact. (4) **CLIP-based Naive Bayes Classifier** (Bayes, 1958) Models prior and conditional probabilities under a conditional independence assumption. Concept polarity is computed as the log-ratio between predicted and contrasting classes, indicating directional contribution.

MLLM Baselines. We evaluate the inherent image classification and reasoning capabilities of MLLMs under two settings. Due to the large label space, which makes it impractical to enumerate all candidate labels in the prompt, we adopt an open-set image classification setup (Zhang et al., 2024a), providing a more challenging and realistic evaluation than the closed-set setup used for weak supervisors. (1) **Zero-shot Input-Output** The model is directly prompted to identify the object in the image without any fine-tuning. (2) **Zero-shot MCoT** The model is prompted to reason step by step. To establish a fair and informative baseline, we augment the prompt "Let's think step by step" (Kojima et al., 2022) with a comprehensive set of diverse concepts, systematically summarized from the concept bank to guide the reasoning process. GPT-4o (OpenAI et al., 2024) is used to quantitatively evaluate the interpretability of the reasoning. To isolate the effect of MCoTs, we further include a fine-tuning-based baseline: (3) **Instruction Tuning (w/o MCoTs)**, The model is fine-tuned on question-answer pairs without explanatory rationales.

Dataset	Weak-Supervision-Guided MCoTs	Dataset	Weak-Supervision-Guided MCoTs
CUB 	**Question** Identify the bird in the image? **Rationale** The image shows that the color of the bird's breast is white; the color of the bird's underparts is white; the shape of the bird's bill is hooked (like a seabird's); the shape of the bird's wings is tapered; the shape of the bird is gull-like. It can be observed that the bird lacks the following features: the pattern on the bird's back is not solid; the color of the bird's under tail is not brown. **Answer**: Laysan Albatross.	LAD-A 	**Question** Identify the animal in the image? **Rationale** The image shows that the color of the animal is black; the teeth are buck-toothed; the diet is leaf-based; the habitat is jungle. **Answer**: Gorilla.
LAD-E 	**Question** Identify the electronic device in the image? **Rationale** The image shows that the size is big (compared to a mobile phone); the visible parts are a motor, a fan and indicator lights; the aim is display; the power rating is low-power. It can be observed that the image lacks the following features: the visible parts are not a plug. **Answer**: Power Supply Unit.	LAD-F 	**Question** Identify the fruit in the image? **Rationale** The image shows that the outside color is red; the epidermis texture is peel-covered; the current state is complete; the hardness is soft; the edibility is common, directly edible and water-rich; the growth pattern is tree-grown; the medicinal property is mild. **Answer**: Plum.
LAD-V 	**Question** Identify the vehicle in the image? **Rationale** The image shows that the color is white; the speed is fast-moving; the parts are a number plate; the aim is rescuing; the price is expensive. It can be observed that the image lacks the following features: the aim is not engineering; the power source is not gasoline. **Answer**: Ambulance.	Oxford-Pets 	**Question** Identify the pet in the image? **Rationale** The image shows that short, smooth coat with a predominantly white color pattern; wide, deep muzzle with a well-defined stop; wide and prominent forehead with a gentle slope. **Answer**: Japanese Chin.

Figure 2: Several examples of generated rationales by WISE.

Implementation. We conduct experiments using Qwen2-VL-7B-Instruct (Wang et al., 2024) as the target MLLM and adopt CLIP ViT-L/14 (Radford et al., 2021) as the backbone of weak supervisors for most datasets, with a parameter count that is approximately 5% of the target model. For SkinCon, we use MAKE (Yan et al., 2025b), pretrained on the million-scale dermatology dataset Derm1M (Yan et al., 2025a), as the backbone. We fine-tune the MLLM using LoRA (Hu et al., 2022) with rank 8 for 10 epochs, employing a total batch size of 16 and a learning rate of 1×10^{-4}. All experiments are conducted on 4 NVIDIA RTX A5000 GPUs.

Evaluation Metrics. For MLLM-based methods, we report classification accuracy and the interpretability. Interpretability is quantified as the proportion of concept polarities in the rationales that agree with expert-annotated binary labels, following the standard CBM evaluation protocol. Since our approach reasons over a subset of concepts, this measure corresponds to concept precision.

4.3 Main Results

Table 1 presents the overall evaluation results of our method on eight concept-annotated datasets.

Interpretability. Weak supervisors effectively guide MLLMs to acquire concept-driven reasoning abilities for image classification, achieving an average interpretability score of 87.83% across eight datasets, which is unexpectedly 8.2% higher than the best-performing weak supervisor (decision trees). On only two datasets does the MLLM's interpretability fall short of its supervisor, further underscoring the central role of decision trees. Compared to the MLLM's inherent zero-shot MCoT, our approach improves interpretability by 37%, consistently outperforming across all datasets.

Accuracy. Differing from most CLIP-based CBMs, which often improve interpretability with slight compromises in accuracy, our method achieves both. Compared to instruction tuning using input-output pairs without MCoTs, our method improves the average accuracy by 0.69%, with only minor drops observed on two datasets. Notably, when the MLLM's inherent reasoning ability is weak, such as on LAD-H and SkinCon-9, where the average interpretability is only 26%, guided reasoning yields substantially larger improvements in classification performance, with gains of up to 2%. These results highlight the critical role of multi-step reasoning in enhancing final decision-making.

Interpretability–Accuracy Trade-Off. A clear trade-off between accuracy and interpretability is observed among weak supervisors: models with higher classification accuracy often exhibit lower interpretability. In contrast, our method integrates the strengths of all weak supervisors when constructing MCoTs, enabling interpretability to generalize effectively from weak to strong as the MCoTs align with human reasoning patterns. However, when the concept-level accuracy of zero-shot reasoning drops to 50.75%, hallucinations tend to arise and propagate to the final predictions, thereby causing an average performance drop of 0.4%.

	Pos	Neg	InCoT	XCoT	Bank
CUB	62.34	69.92	7	301	312
SkinCon-3	66.01	89.16	17	22	22
SkinCon-9	75.00	94.41	23	48	48
LAD-A	82.78	-	3	110	123
LAD-E	98.62	100.00	3	72	75
LAD-H	86.86	73.81	4	22	22
LAD-F	97.29	89.47	5	53	58
LAD-V	95.15	94.73	3	76	81
Average	83.01	87.36	8	88	93

Table 2: **Analysis of MCoTs**. *Pos* and *Neg* denote the precision of supportive and refutational concepts of the MLLM-generated MCoTs on the test set, respectively. *InCoT* indicates the average number of concepts used per image for reasoning, *XCoT* denotes the number of unique concepts used at least once across all images, and *Bank* refers to the total number of expert-defined concepts provided for weak supervision.

Method	Overall	Pos-C	Neg-C
Ours	**65.03**	**62.34**	**69.92**
- w/o Salience	32.73	32.65	48.39
- w/o Order	61.46	56.22	69.13
Captioning	49.68	49.68	-
Instance Tree	56.13	52.91	65.99
Category Tree	62.21	56.84	67.74

Table 3: **Ablation studies** on the CUB test set, assessing the effects of different components and variants on the concept-level precision (interpretability) of MCoTs.

Datasets w/o Concept Annotation. We further evaluate WISE on Oxford-Pets (37 classes) and Oxford-Flowers (102 classes), two datasets without concept annotations. Without MCoTs, instruction tuning achieves accuracies of 94.19% and 98.72%, respectively. Incorporating WISE yields 93.87% and 98.88%, indicating comparable performance. Moreover, the case studies in Figure 2 demonstrate that WISE substantially enhances interpretability.

4.4 MCoT Analysis

Table 2 reports the decoupled precision of positive and negative concepts within the MCoTs on the test dataset, representing supportive reasoning and refutational elimination, respectively. In addition, the table presents the number of concepts used in the MCoTs and compares them with the total number of predefined concepts in the bank, highlighting the efficiency and dynamic selectivity of the reasoning.

Concept Contribution. The model achieves an average precision of 83% for positive concepts and 87% for negative concepts, indicating a balanced adoption of both reasoning strategies. These results suggest that the MLLM integrates both supportive and refutational reasoning without exhibiting performance bias. In the relatively simple LAD-A, the decisions can be made solely through supportive reasoning, with no reliance on refutational cues.

Concept Efficiency. Our MCoTs require only 8 concepts on average to complete reasoning across all datasets, consistent with the human intuition that a small number of concepts is sufficient for object identification. In contrast, traditional CBMs score all 312 concepts in the concept bank to reach a decision, which clearly contradicts this intuition. Although recent efforts (Yan et al., 2023a) have reduced this number to 32 on CUB, the inherent scoring mechanism of CBM still leads to redundancy. Notably, our method achieves accurate classification of 200 bird species using only 7 concepts. This efficiency stems from the decision tree's ability to capture logical dependencies among concepts. We also observe that the SkinCon dataset requires significantly more concepts. Manual inspection reveals that this is due to limitations of the concept bank: Some diseases share identical concept patterns, rendering them indistinguishable under the current representation. Future work should consider expanding the concept bank of SkinCon to improve its coverage and discriminative capacity.

Dynamic Concept Selection. We observe that although reasoning over a single image involves only 8 concepts on average, the MCoTs generated by WISE dynamically select concepts, resulting in 95% of the concept bank being used at least once. This demonstrates a clear contrast to prior work (Jiang et al., 2025), which relies on a fixed concept bank and scores each concept independently.

4.5 Ablation Study

Table 3 compares the concept-level precision of the generated MCoTs on the CUB test dataset.

Salient Concept Selection. Selecting visually salient concepts substantially improves MLLMs' reasoning precision. CLIP's vision-language alignment scores not only compensate for weak supervisors' limited visual perception, but also reduce MLLMs' reliance on imperceptible cues.

Concept Organization. We create a variant by shuffling the order of concepts in the generated MCoTs, resulting in rationales that are unordered and unstructured. This randomness increases the learning difficulty for the model. In contrast, we introduce human-aligned reasoning biases, reducing such uncertainty and improving concept precision.

Captioning as MCoT. To differentiate concept-driven reasoning from simple image captioning, we train MLLMs using all visual concepts present in the image, arranged in a predefined order. Result indicates caption-style MCoTs fail to facilitate the acquisition of concept-level reasoning.

Variant Tree Construction. We design two tree variants: category-specific and instance-specific. The former builds one tree per category, ignoring individual variability, whereas the latter relies on instance-level feature saliency, neglecting cross-category regularities. Consequently, both methods suffer reduced precision due to their lack of shared structure and sensitivity to individual differences.

5 Discussion

We demonstrated the efficiency and dynamic nature of concept selection in Section 4.4. Beyond these observations, WISE offers two notable advantages:

Stepwise Transparent Reasoning. Guided by decision trees, WISE generates MCoTs by selecting the most discriminative concept at each step based on information gain, progressively isolating the target class from negatives. Each selection explicitly conveys three elements to enhance transparency: (1) the rationale for the choice, (2) the negative classes excluded by this step, and (3) the remaining negatives contributing to non-zero Gini impurity. This stepwise process naturally aligns with the autoregressive behavior of MLLMs.

Quantified Sufficiency. Reasoning continues until a leaf node is reached, where the sufficiency of the reasoning chain is evaluated. A zero Gini impurity indicates complete reasoning, whereas a non-zero impurity (when no further splits are possible) reveals gaps in the current concept bank.

Overall, WISE achieves more efficient and concise concept usage than CLIP-based CBMs such as LM4CV (Yan et al., 2023a) and neural-symbolic reasoning methods like Deep Concept Reasoner (DCR) (Barbiero et al., 2023), while offering distinctive advantages. Unlike DCR's permutation-invariant logical rules, WISE adopts a human-aligned reasoning paradigm that preserves the uniqueness of the generated MCoTs.

6 Future Work

Beyond enabling interpretable image classification with MLLMs, our work lays the foundation for more efficient human-in-the-loop frameworks (Yan et al., 2023b). Concept-based models allow human intervention at the concept level, providing direct control over model behavior (Koh et al., 2020). By generating the minimal set of concepts needed for explanation, our method significantly improves intervention efficiency. This property is particularly valuable in domains such as clinical image diagnosis, where rapid and precise feedback is critical.

In addition, we introduce a new class of instruction data designed to promote intra-object understanding during MLLM pretraining. This resource can support the creation of large-scale MCoT datasets for both domain-specific and general applications. While our current experiments focus on a single dataset, future work should explore dataset integration and scaling toward training foundation-level MLLMs (Zhang et al., 2024a). Finally, the released MCoT dataset provides a practical benchmark for diverse vision tasks, including hallucination evaluation and mitigation (Bai et al., 2025), enabling systematic comparison of hallucination-reduction strategies and advancing reliable MLLM reasoning.

7 Conclusion

We propose WISE, a method that reformulates the concept bottleneck layer into natural-language MCoTs guided by weak supervision, aligning the model's reasoning with human thought patterns. This method is broadly applicable to any image classification dataset with category labels. Experiments show that the generated MCoTs yield a 37% improvement in the interpretability of MLLMs.

Limitations

A main limitation of our approach is its dependence on concept banks generated by prompting LLMs for datasets lacking predefined annotations. The quality of the resulting MCoTs is tied to the coverage of these banks, and insufficient concept sets may reduce the model's ability to distinguish visually similar categories.

Moreover, although we reformulate image classification as a question answering task, most datasets are designed with a fixed and limited label space. This mismatch can introduce concept conflicts when integrating multiple MCoT-augmented datasets. For example, generated rationales may fail to eliminate distractor labels that are absent from one dataset but present in others. These issues highlight the need for comprehensive datasets that adequately represent the target domain.

References

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2025. Hallucination of multimodal large language models: A survey. *Preprint*, arXiv:2404.18930.

Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, Alberto Tonda, Pietro Lio, Frederic Precioso, Mateja Jamnik, and Giuseppe Marra. 2023. Interpretable neural-symbolic concept reasoning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 1801–1825. PMLR.

Thomas Bayes. 1958. An essay towards solving a problem in the doctrine of chances. *Biometrika*, 45(3-4):296–315.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees*. Chapman and Hall/CRC, New York.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. 2024. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4971–5012. PMLR.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024. M^3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221, Bangkok, Thailand. Association for Computational Linguistics.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Hao Zhang, and Chuang Gan. 2023. See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning. *Preprint*, arXiv:2301.05226.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. 2025. CoMT: A novel benchmark for chain of multi-modal thought on large vision-language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23678–23686.

Vinícius G. Costa and Carlos Eduardo Pedreira. 2023. Recent advances in decision trees: An updated survey. *Artificial Intelligence Review*, 56(5):4765–4800.

Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto Novoa, and James Y Zou. 2022. SkinCon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. In *Advances in Neural Information Processing Systems*, volume 35, pages 18157–18167. Curran Associates, Inc.

Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, and Rongrong Ji. 2024. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 9096–9105, New York, USA. Association for Computing Machinery.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*, 2nd edition. Springer, New York.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yiwen Jiang, Deval Mehta, Wei Feng, and Zongyuan Ge. 2025. Enhancing interpretable image classification through LLM agents and conditional concept bottleneck models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12285–12297, Vienna, Austria. Association for Computational Linguistics.

Yiwen Jiang, Hao Yu, and Xingyue Fu. 2023. Medical decision tree extraction: A prompt based dual contrastive learning method. In *Health Information Processing. Evaluation Track Papers*, pages 103–116, Singapore. Springer Nature Singapore.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2673–2682. PMLR.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy

Liang. 2020. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Max Losch, Mario Fritz, and Bernt Schiele. 2019. Interpretability beyond classification output: Semantic bottleneck networks. *Preprint*, arXiv:1907.10882.

Xuewen Luo, Fan Ding, Yinsheng Song, Xiaofeng Zhang, and Junnyong Loo. 2025. PKRD-CoT: A unified chain-of-thought prompting for multi-modal large language models in autonomous driving. In *Neural Information Processing*, pages 62–76, Singapore. Springer Nature Singapore.

Deval Mehta, Yiwen Jiang, Catherine Jan, Mingguang He, Kshitij Jadhav, and Zongyuan Ge. 2025. Interpretable few-shot retinal disease diagnosis with concept-guided prompting of vision-language models. In *Information Processing in Medical Imaging*, pages 263–277, Cham. Springer Nature Switzerland.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *Proceedings of the Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, page 722–729, USA. IEEE Computer Society.

Tuomas P. Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. 2023. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Winnie Pang, Xueyi Ke, Satoshi Tsutsui, and Bihan Wen. 2024. Integrating clinical knowledge into concept bottleneck models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 243–253, Cham. Springer Nature Switzerland.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2017. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, page 506–519, New York, USA. Association for Computing Machinery.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE Computer Society.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Chenming Shang, Shiji Zhou, Hengyuan Zhang, Xinzhe Ni, Yujiu Yang, and Yuwang Wang. 2024. Incremental residual concept bottleneck models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 11030–11040. IEEE.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. California Institute of Technology.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian J. McAuley. 2023a. Learning concise and descriptive attributes for visual recognition. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3067–3077. IEEE.

Siyuan Yan, Ming Hu, Yiwen Jiang, Xieji Li, Hao Fei, Philipp Tschandl, Harald Kittler, and Zongyuan Ge. 2025a. Derm1m: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. *Preprint*, arXiv:2503.14911.

Siyuan Yan, Xieji Li, Ming Hu, Yiwen Jiang, Zhen Yu, and Zongyuan Ge. 2025b. Make: Multi-aspect knowledge-enhanced vision-language pretraining for zero-shot dermatological assessment. *Preprint*, arXiv:2505.09372.

Siyuan Yan, Zhen Yu, Xuelin Zhang, Dwarikanath Mahapatra, Shekhar S. Chandra, Monika Janda, H. Peter Soyer, and Zongyuan Ge. 2023b. Towards trustable skin cancer diagnosis via rewriting model's decision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 11568–11577. IEEE.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. 2023. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19187–19197. IEEE.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Mert Yüksekgönül, Maggie Wang, and James Zou. 2023. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. 2024a. Why are visually-grounded language models bad at image classification? In *Advances in Neural Information Processing Systems*, volume 37, pages 51727–51753. Curran Associates, Inc.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024b. Multimodal chain-of-thought reasoning in language models. *Preprint*, arXiv:2302.00923.

Bo Zhao, Yanwei Fu, Rui Liang, Jiahong Wu, Yonggang Wang, and Yizhou Wang. 2019. A large-scale attribute dataset for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Haojie Zheng, Tianyang Xu, Hanchi Sun, Shu Pu, Ruoxi Chen, and Lichao Sun. 2024. Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination. *Preprint*, arXiv:2411.12591.

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Interpretable basis decomposition for visual explanation. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII*, page 122–138, Berlin, Heidelberg. Springer-Verlag.