Stable Vision Concept Transformers for Medical Diagnosis

Lijie Hu[*,1,2], Songning Lai[*,1,2], Yuan Hua[*,1,2,3], Shu Yang[1,2], Jingfeng Zhang[1,2,4], Di Wang[†,1,2]

[1] Provable Responsible AI and Data Analytics (PRADA) Lab
[2] King Abdullah University of Science and Technology
[3] Tsinghua University
[4] University of Auckland

Abstract. Transparency is a paramount concern in the medical field, prompting researchers to delve into the realm of explainable AI (XAI). Among these XAI methods, Concept Bottleneck Models (CBMs) aim to restrict the model's latent space to human-understandable high-level concepts by generating a conceptual layer for extracting conceptual features, which has drawn much attention recently. However, existing methods rely solely on concept features to determine the model's predictions, which overlook the intrinsic feature embeddings within medical images. To address this utility gap between the original models and concept-based models, we propose **V**ision **C**oncept **T**ransformer (**VCT**). Furthermore, despite their benefits, CBMs have been found to negatively impact model performance and fail to provide stable explanations when faced with input perturbations, which limits their application in the medical field. To address this faithfulness issue, this paper further proposes the **S**table **V**ision **C**oncept **T**ransformer (**SVCT**) based on VCT, which leverages the vision transformer (ViT) as its backbone and incorporates a conceptual layer. SVCT employs conceptual features to enhance decision-making capabilities by fusing them with image features and ensures model faithfulness through the integration of Denoised Diffusion Smoothing. Comprehensive experiments on four medical datasets demonstrate that our VCT and SVCT maintain accuracy while remaining interpretability compared to baselines. Furthermore, even when subjected to perturbations, our SVCT model consistently provides faithful explanations, thus meeting the needs of the medical field.

Keywords: Explainable medical image classification · Explainability · Stability · Medical diagnosis.

1 Introduction

As the field of medical image analysis continues to evolve, deep learning models and methods have demonstrated excellent performance in tasks such as image

[*] Equal Contribution.
[†] Corresponding Author.

recognition and disease diagnosis [9]. However, these advanced deep learning models are usually regarded as black boxes and lack credibility and transparency. Especially in the medical field, this opacity makes it difficult for physicians and clinical professionals to trust the predictions of the models. Thus, the requirement for interpretability of model decisions is more urgent in the medical field [17].

The healthcare field, characterized by stringent requirements for trustworthiness, necessitates models that not only exhibit high performance but are also comprehensible and can be trusted by practitioners. Therefore, Explainable Artificial Intelligence (XAI) has become one of the hotspots for research and development. By introducing interpretability, XAI tries to make the decision-making process of deep learning models more transparent and understandable. Some compelling interpretable methods, such as attention mechanisms [20], saliency maps [26], DeepLIFT and Shapley values [12], and influence functions [10], attempt to provide users with visual explanations about model decisions. However, while these post-hoc explanatory methods can provide useful information, there is still a certain disconnect between their explanations and model decisions, and these explanations are generated after model training and fail to participate in the model learning process. Some studies [17] have shown that post-hoc is sensitive to slight changes in the input, making the post-hoc methods misleading as they could provide explanations that do not accurately reflect the model's decision-making process.

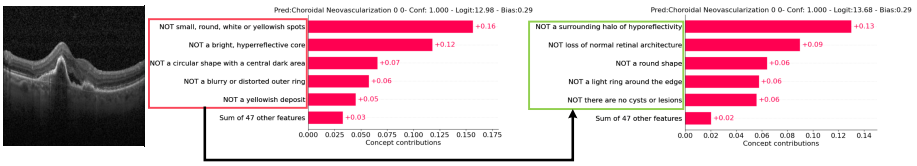

Fig. 1: An example of VCT framework on OCT2017 dataset [9]. The leftmost figure displays the input image, while the adjacent one on the left shows the concept output without perturbations. In contrast, the figure on the right presents the concept output after applying input perturbations, resulting in noticeable changes.

Therefore, researchers have shown interest in self-explained methods. Among them, concept-based methods have attracted a lot of attention. These approaches strive to incorporate interpretability into machine learning models by establishing connections between their predictions and concepts that are understandable to humans. As an illustration, the Concept Bottleneck Model (CBM) [11] initially forecasts an intermediate set of predefined concepts, subsequently utilizing these concepts to make predictions for the final output. [15] introduce Label-free CBM, a novel framework designed to convert any neural network into an interpretable CBM without the need for labeled concept data compared to the original CBM. These inherently interpretable methods provide concept-based explanations, which are generally more comprehensible than post-hoc approaches.

However, many existing methods rely solely on concept features to determine the model's predictions. These approaches overlook the intrinsic feature embeddings within medical images. For instance, [18] solely utilizes concept labels to supervise the concept prediction results of the entire image. This oversight can lead to a decrease in classification accuracy, which is suggested to stem from the inefficient utilization of valuable medical information. Therefore, a significant challenge in the field of medical imaging is how to maintain a high level of accuracy while incorporating interpretability.

To address the aforementioned challenges, we propose Vision Concept Transformer (VCT), a novel medical image processing framework that is interpretable and maintains high performance. Vision Transformers (ViTs) [3] have achieved state-of-the-art performance for various vision tasks, showing good robustness in prediction. Thus, in the VCT framework, we utilize ViTs as the foundational network. To enhance interpretability, we employ a label-free methodology for generating the conceptual layer. Moreover, unlike previous CBMs, which only use conceptual features for prediction, in the VCT framework, we integrate conceptual features with image features, utilizing the conceptual layer as supplementary information to augment decision-making. This integration effectively addresses the issue of accuracy degradation associated with a singular label-free CBM, ensuring interpretability without compromising accuracy.

While VCT keeps the interpretability of CBMs, it also inherits their interpretability instability when facing perturbations or noise in the input. Specifically, adding slight noise to the input image can significantly change the top-k important concepts given by CBMs (see Figure 1 for an example), i.e., the top k-indices of the concept vector. Instability is a common issue in deep learning interpretation methods, making it challenging to understand model reasoning [6], especially with unlabeled data and self-supervised training [4]. As in real medical scenarios, there is always natural and inherent noise or some adversarial examples manipulated by attackers [1]. Thus, VCT cannot be a faithful explainable tool for these applications.

To address the faithfulness issue, by using the Denoised Diffusion Smoothing method, we can smoothly and directly transform VCT into a Stable Vision Concept Transformer (SVCT) framework that is capable of providing stable interpretations despite perturbations to the inputs, the structure is shown in Figure 2. Our contributions can be summarised as follows.

- We proposed the VCT framework, transforming ViTs into an interpretable CBM. VCT integrates conceptual features with image features, utilizing conceptual features as auxiliary decision-making components. This effectively addresses the performance degradation issue in existing CBMs due to inefficient utilization of medical information.
- To further enhance the interpretability stability of VCT, we propose a formal mathematical definition of an SVCT, which ensures that the top-k index of its conceptual vectors remains relatively stable under slight perturbations. We utilize a Denoised Diffusion Smoothing (DDS) method to obtain

an SVCT. Moreover, we theoretically proved that our method satisfies the properties of SVCT.
– We conducted extensive experiments on four medical datasets to validate the superiority of SVCT in the medical domain. First, we demonstrate that our SVCT is more accurate and interpretable than other CBM approaches. Secondly, we verified that the SVCT model still provides stable explanations under perturbations.

2 Related Work

Concept Bottleneck Models. Concept Bottleneck Model (CBM) [11] stands out as an innovative deep-learning approach applied to image classification and visual reasoning. It introduces a concept bottleneck layer into deep neural networks, enhancing model generalization and interpretability by learning specific concepts. However, CBM faces two primary challenges: its performance often lags behind that of original models lacking the concept bottleneck layer, attributed to incomplete information extraction from the original data to bottleneck features. Additionally, CBM relies on laborious dataset annotation [7]. Researchers have explored solutions to these challenges. [2] extend CBM into interactive prediction settings, introducing an interaction policy to determine which concepts to label, thereby improving final predictions. [14] address CBM limitations and propose a novel framework called Label-free CBM. This innovative approach enables the transformation of any neural network into an interpretable CBM without requiring labeled concept data, all while maintaining high accuracy [24]. However, most of the existing CBMs use only conceptual features for prediction, which can cause a degradation in prediction performance and make them unsuitable for medical scenarios.

Faithfulness in Explainable Methods. Faithfulness is an important property that should be satisfied by explanatory models, which ensures that the explanation accurately reflects the true reasoning process of the model [8]. Stability is crucial to the faithfulness of the interpretation. Some preliminary work has been proposed to obtain stable interpretations. For example, [23] theoretically analyzed the stability of post-hoc explanations and proposed the use of smoothing to improve the stability of explanations. They devised an iterative gradient descent algorithm for obtaining counterfactual explanations, which showed desirable stability. However, these techniques are designed for post-hoc explanations and cannot be directly applied to attention-based mechanisms like ViTs.

Interpretability in Medical Image Classification. In the research of interpretable artificial intelligence in medical image analysis, [21] proposes a new method to construct a robust and interpretable medical image classifier using natural language concepts, and it has been evaluated on multiple datasets. [18] focuses on self-explanatory deep models, introducing a model that implicitly learns conceptual explanations during training by adding an explanation generation module.

These methods collectively enhance the interpretability of the model. However, the existing interpretability methods face two main issues. Firstly, they rely solely on concept features for decision-making, leading to insufficient utilization of valuable information in medical images and resulting in a performance decline in medical image processing. Secondly, existing methods exhibit instability when confronted with noise, failing to provide faithful explanations. Therefore, our work aims to ensure good performance while maintaining interpretability and providing faithful explanations to address these issues. See Appendix F for more details.

3 Stable Vision Concept Transformer

In this section, we propose the Stable Vision Concept Transformer (SVCT) framework. Specifically, we first leverage the Label-free Concept Bottleneck Model [15] to transform the ViT network into an interpretable CBM without concept labels, which is an automated, scalable, and efficient fashion to address the core limitations of existing CBMs. We then fuse the concept features with the ViTs features as decision-aiding features, which not only improves the interpretability of the model but also ensures a high degree of accuracy. To obtain an SVCT, we adopt Denoised Diffusion Smoothing (DDS) to turn it into an SVCT.

Our model consists of the following six steps, which are illustrated in Figure 2 - **Step1:** The ViT model is trained on the target task, and VCT is transformed into SVCT by inserting the DDS method. **Step2:** We generate initial concept set based on the target task and filter out unwanted concepts using a series of filters. **Step3:** Compute embeddings by the backbone on the training dataset and obtain the concept matrix. **Step4:** Learn projection weights W_c to create a Concept Bottleneck Layer (CBL). **Step5:** Fuse the concept features with the ViTs features. **Step6:** Learn the weights W_F of the sparse final layer to make predictions. Detailed notations can be found in Table 6. We first introduce VCT for convenience.

3.1 Vision Concept Transformer

In this section, we introduce the vision concept transformer. Before that, it is necessary to pre-train the ViT model f on the target task dataset as a backbone for the VCT framework.

Label-free CBMs. We use the label-free CBM [15] to get concept feature $f_c(X) \in \mathbb{R}^M$, where M is the number of concepts. Firstly, we obtain a concept set and use it as human-understandable concepts in the concept bottleneck layer (See Appendix D and E for details). Next, we need to learn how to project from the feature space $\mathbb{R}^{d_0}$ of the backbone network to an interpretable feature space $\in \mathbb{R}^M$ that corresponds to the set of interpretable concepts in the axial direction. We use a way of learning the projection weights $W_c \in \mathbb{R}^{M \times d_0}$ without any labeled concept data by utilizing CLIP-Dissect [16]. We can learn about a bottleneck

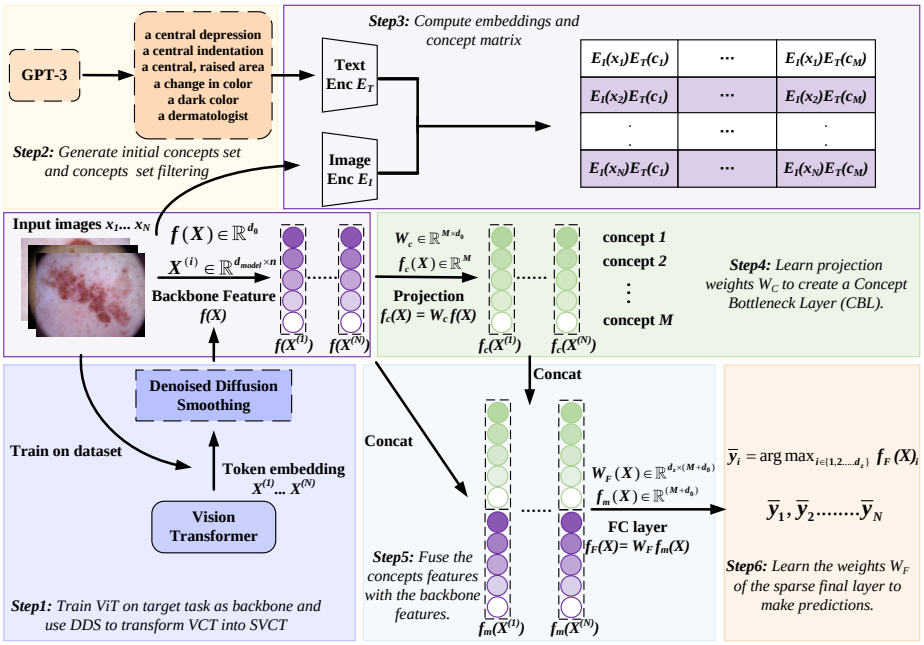

Fig. 2: Overview of our Stable Vision Concept Transformer (SVCT) model.

conceptual layer and get the concept feature

$$f_c(X) = W_c f(X) \in \mathbb{R}^M. \tag{1}$$

Concat ViT feature and concept feature. Now that we have learned about the conceptual bottleneck layer and get $W_c \in \mathbb{R}^{M \times d_0}$. In VCT, the conceptual features are no longer used as the only features for classification. According to previous studies, based on the conceptual features alone will degrade the accuracy of the model. Therefore, here we use the conceptual features as the supplementary features, which are fused with the features extracted from the backbone network, and this feature fusion makes the VCT able to ensure accuracy improvement while having a better explanatory nature. Specifically, we define $f_m(X) = \text{concat}(f(X), f_c(X))$, where $f_m\left(X^{(i)}\right) \in \mathbb{R}^{M+d_0}$, and we define a feature of VCTs for prediction as follows:

$$F(X) = \text{concat}(f(X), W_c f(X)). \tag{2}$$

Final classification layer. The next goal is to learn the final predictor using the fully connected layer $W_F \in \mathbb{R}^{d_z \times (M+d_0)}$, where d_z represents the final number of predicted categories. For each input X, we have access to its predictive distribution through the final classification layer.

3.2 Stable VCT

As we mentioned in the introduction and Figure 1, CBMs and VCT have an interpretation instability issue, i.e., a slight perturbation on the input could

change the top-k concepts in the concept vector (concept feature in VCT). Here we aim to address the instability issue. We first give the definition of the top-k overlap ratio for two (concept) vectors,

Definition 1. *For vector $x \in \mathbb{R}^n$, we define the set of top-k component $T_k(\cdot)$ as*

$$T_k(x) = \{i : i \in [d] \text{ and } \{|\{x_j \geq x_i : j \in [n]\}| \leq k\}\}.$$

For two vectors x, x', their top-k overlap ratio $V_k(x, x')$ is defined as $V_k(x, x') = \frac{1}{k}|T_k(x) \cap T_k(x')|$.

Definition 2 (Stable VCTs). *Giving M number of concepts, a norm $\|\cdot\|$, and a divergence metric D, we call a function $g : \mathbb{R}^{d_{model} \times n} \to \mathbb{R}^M$ is an $(R, D, \gamma, \beta, k, \|\cdot\|)$-stable concept module for VCTs if for any given input data X and for all $X' \in \mathbb{R}^{d_{model} \times n}$ such that $\|X - X'\| \leq R$:*

(1) (Explanation Stability) $V_k\left(g\left(X'\right), g(X)\right) \geq \beta$.
(2) (Prediction Robustness) $D\left(\bar{y}(X), \bar{y}\left(X'\right)\right) \leq \gamma$, where $\bar{y}(X), \bar{y}\left(X'\right)$ are the prediction distribution of VCTs based on $g(X), g\left(X'\right)$ respectively.

We call the models of VCTs based on g as SVCTs.

Intuitively, for input X, $g(X)$ is its concept vector. Thus, the first condition of SVCT ensures that the k-most important concepts will not change much, even if there are some perturbations on the input. The second one guarantees that the prediction of SVCT is also stable against perturbation, which inherits the good performance of VCT. For the parameters, R represents the stable radius. Within this radius, g is a stable concept module, D is the Rényi divergence between two distributions (we denote it as D_α). γ is a similarity coefficient, and as γ gets smaller, g is more robust. β is the stability coefficient, which measures the stability of the interpretation, and as β gets larger, g is more stable. In this paper, $\|\cdot\|$ is the ℓ_2-norm (if we consider X as a $d = d_{model} \times n$ dimensional vector). We can show if the prediction distribution is robust under Rényi divergence, then the prediction will be unchanged with perturbations on input (shown in Theorem 1).

Theorem 1. *If a function is a $(R, D_\alpha, \gamma, \beta, k, \|\cdot\|)$-stable concept module for VCTs, then if*

$$\gamma \leq -\log(1 - p_{(1)} - p_{(2)} + 2(\frac{1}{2}(p_{(1)}^{1-\alpha} + p_{(2)}^{1-\alpha}))^{\frac{1}{1-\alpha}}),$$

we have for all X' such that where $\|X - X'\| \leq R$,

$$\arg\max_{h \in \mathcal{H}} \mathbb{P}(\bar{y}(X) = h) = \arg\max_{h \in \mathcal{H}} \mathbb{P}(\bar{y}(X') = h),$$

where $\mathcal{H}$ is the set of classes, $p_{(1)}$ and $p_{(2)}$ refer to the largest and the second largest probabilities in $\{p_i\}$, where p_i is the probability that $\bar{y}(X)$ returns the i-th class.

Finding Stable Vision Concept Transformers. Motivated by [5], we propose a method called Denoised Diffusion Smoothing (DDS) to obtain SVCTs. The process is as follows: we use randomized smoothing to the VCT and then apply a denoised diffusion probabilistic model to the perturbed input. With this processing, we can transform a VCT into an SVCT, and its corresponding concept module becomes a stable concept module. Specifically, for a given input image x, its corresponding token embedding is X. We add some randomized Gaussian noise to X, i.e., $\tilde{X} = X + S$, where $S \sim \mathcal{N}\left(0, \sigma^2 I_{d_{model} \times n}\right)$. Then we will use some denoised diffusion models to denoise $\tilde{X}$ to get $\hat{X}$. We then take the obtained $\hat{X}$ as a new input to get concept feature $f_c(\hat{X})$ in (1) and go through the remaining structures of the VCT to get the final prediction.

Specifically, for a given input X, randomized smoothing is done by augmenting the data points of an image by adding additive Gaussian noise to the image, which we can denote as $X_{\text{rs}} \sim \mathcal{N}\left(X, \sigma^2 \mathbf{I}\right)$. Diffusion models rely on a particular form of noise modeling, denoted as $X_t \sim \mathcal{N}\left(\sqrt{\beta_t} X, (1-\beta_t)\,\mathrm{I}\right)$. Where β_t is a constant related to time step t. Thus, if we want to use a diffusion model for randomized smoothing, we need to establish a link between the parameters of the two noise models. The DDS model used in this paper multiplies X_{rs} by the factor $\sqrt{\beta_t}$, thus satisfying the requirement of the noise mean, and accordingly, in order to satisfy the requirement of the variance, we can obtain the equation $\sigma^2 = \frac{1-\beta_t}{\beta_t}$. As the time step changes, σ^2 changes as β_t changes because β_t is a constant with respect to the time step. But it can be computed at every time step, and by using this, we are able to obtain $X_{t^*} = \sqrt{\beta_{t^*}}(X + S)$, where $S \sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}\right)$. Such a form of noise is consistent with the form on which the diffusion model depends, and we can use the diffusion model on X_{t^*} to obtain denoised sample $\hat{X} = \text{denoise}\left(X_{t^*}; t^*\right)$. In this paper, we repeat this process several times to improve robustness.

In the following, we show that $\tilde{w} = f_c(\hat{X})$ is a stable concept feature satisfying Definition 2 if σ^2 satisfies some condition. Before showing the results, we first provide some notations. For input image x, we denote $\tilde{w}_{i^*}$ as the i-th largest component in $\tilde{w}(x)$. Let $k_0 = \lfloor (1-\beta)k \rfloor + 1$ as the minimum number of changes on $\tilde{w}(x)$ to make it violet the β-top-k overlapping ratio with $\tilde{w}(x)$. Let $\mathcal{S}$ denote the set of last k_0 components in top-k indices and the top k_0 components out of top-k indices. Then, we can prove the following upper bound. The details of the algorithm are in Algorithm 1.

Algorithm 1 SVCTs via Denoised Diffusion Smoothing

1: **Input:** X; A standard deviation $\sigma > 0$.
2: t^*, find t s.t. $\frac{1-\beta_t}{\beta_t} = \sigma^2$.
3: $X_{t^*} = \sqrt{\beta_{t^*}}(\tilde{X} + \mathcal{N}(0, \sigma^2 \mathbf{I}))$.
4: $\hat{X} = \text{denoise}(X_{t^*}; t^*)$.
5: $w = f_c(\hat{X})$, where f_c is in (1).
6: **Return:** Concept feature vector w.

Theorem 2. *Consider the function* $\tilde{w}(X) = f_c(T(X+S))$*, where* f_c *as the function in (1),* T *as the denoised diffusion model and* $S \sim \mathcal{N}(0, \sigma^2 I_{d_{model} \times n})$*. Then, it is an* $(R, D_\alpha, \gamma, \beta, k, \|\cdot\|_2)$*-stable concept module for VCTs for any* $\alpha > 1$ *if for any input image* x *we have*

$$\sigma^2 \geq \max\{\alpha R^2/2(\frac{\alpha}{\alpha-1}\ln(2k_0(\sum_{i\in\mathcal{S}}\tilde{w}_{i^*}^{\alpha})^{\frac{1}{\alpha}} + (2k_0)^{\frac{1}{\alpha}}\sum_{i\notin\mathcal{S}}\tilde{w}_{i^*}) - \frac{1}{\alpha-1}\ln(2k_0)), \alpha R^2/2\gamma\}.$$

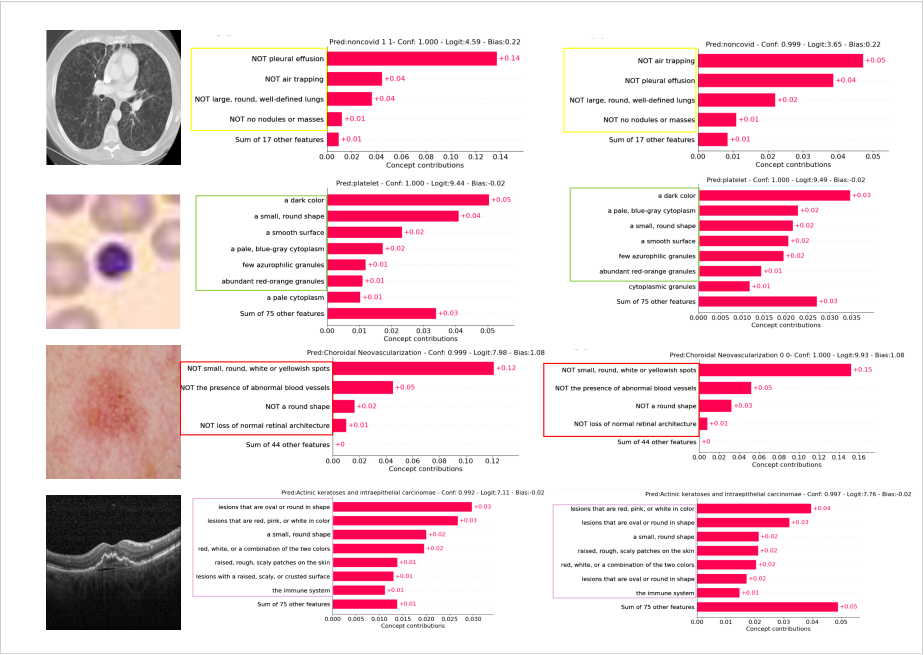

Fig. 3: Results of concept visualization. From left to right: one sample from each dataset, concept visualization results before perturbation, and concept visualization results after perturbation. Clear and enlarged pictures are shown in the Appendix L.

4 Experiments

4.1 Experimental Settings

Datasets. We conducted experiments on four medical datasets, including Human Against Machine with 10,015 training images (HAM10000) dataset [19],

Covid19-CT dataset [25], BloodMNIST dataset [22], and Optical coherence tomography (OCT) 2,017 dataset [9]. Details are in Appendix G.

Table 1: Results of accuracy for the baselines and SVCT w/w.o perturbation.

Method	**HAM10000**	**Covid19-CT**	**BloodMNIST**	**OCT2017**
Standard (No interpretability)	99.13%	81.62%	97.05%	99.70%
Label-Free CBM (LF-CBM)	93.61%	79.75%	94.97%	97.50%
Post-hoc CBM (P-CBM)	97.60%	76.26%	94.83%	98.60%
Vision Concept Transformer (VCT)	99.00%	80.62%	96.21%	99.10%
Stable VCT(SVCT)	**99.05%**	**81.37%**	**96.96%**	**99.50%**
$\rho_u = 8/255$ - LF-CBM	90.08%	67.98%	80.53%	91.88%
$\rho_u = 8/255$ - P-CBM	90.96%	70.66%	77.55%	91.70%
$\rho_u = 8/255$ - VCT	95.80%	69.78%	89.45%	96.80%
$\rho_u = 8/255$ - **SVCT**	**97.97%**	**74.45%**	**94.07%**	**98.70%**
$\rho_u = 10/255$ - LF-CBM	88.70%	65.12%	75.63%	90.58%
$\rho_u = 10/255$ - P-CBM	90.21%	66.32%	74.27%	90.10%
$\rho_u = 10/255$ - VCT	95.28%	68.85%	87.71%	96.25%
$\rho_u = 10/255$ - **SVCT**	**97.24%**	**71.65%**	**92.65%**	**98.48%**

Baselines. In this paper, the standard model is ViT [3], which accomplishes the classification task by extracting image features, but the model itself is not interpretable. The baseline model is label-free CBM [15], which uses ViT as the backbone to generate a conceptual bottleneck layer and finally makes predictions through a linear layer.

Table 2: Results on CFS and CPCS for the baselines and SVCT under various perturbations.

Method	**HAM10000**		**Covid19-CT**		**BloodMNIST**		**OCT2017**	
	CFS	CPCS	CFS	CPCS	CFS	CPCS	CFS	CPCS
$\rho_u = 6/255$ - LF-CBM	0.3335	0.9405	0.6022	0.8117	0.5328	0.8511	0.3798	0.9254
$\rho_u = 6/255$ - VCT	0.3361	0.9394	0.6761	0.7650	0.5432	0.8436	0.3625	0.9314
$\rho_u = 6/255$ - **SVCT**	**0.1354**	**0.9900**	**0.5555**	**0.8359**	**0.3589**	**0.9320**	**0.3257**	**0.9468**
$\rho_u = 8/255$ - LF-CBM	0.3719	0.9256	0.6707	0.7710	0.6280	0.7947	0.3941	0.9196
$\rho_u = 8/255$ - VCT	0.4109	0.9098	0.8114	0.6743	0.7162	0.7328	0.3812	0.9240
$\rho_u = 8/255$ - **SVCT**	**0.1555**	**0.9867**	**0.6446**	**0.7818**	**0.4383**	**0.8977**	**0.3459**	**0.9387**
$\rho_u = 10/255$ - LF-CBM	0.4027	0.9123	0.7224	0.7336	0.6906	0.7545	0.4055	0.9145
$\rho_u = 10/255$ - VCT	0.4637	0.8844	0.8943	0.6155	0.8057	0.6670	0.3949	0.9179
$\rho_u = 10/255$ - **SVCT**	**0.1725**	**0.9836**	**0.7096**	**0.7389**	**0.5058**	**0.8625**	**0.3620**	**0.9321**

Perturbations. Perturbation refers to small changes or modifications made to input data. In this paper, we introduce perturbations to input images with different radius ρ_u to assess the stability and robustness of the SVCT model.

The range of perturbation radii ρ_u is [6/255, 10/255]. We employ the PGD [13] algorithm to craft adversarial examples with a step size of 2/255 and a total of 10 steps. As a default, we set the standard deviation $S = 8/255$ for the Gaussian noise in our method. All results are the average score running 10 times to reduce variance.

Evaluation metrics. To demonstrate the utility of our approach, we report the classification accuracy on test data for classification tasks. We evaluate our model's stability using Concept Faithfulness Score (CFS) and Concept Perturbation Cosine Similarity (CPCS). CFS measures the stability of model interpretability between two concept weight vectors using Euclidean distance; we use c_1 to represent the concept weight vector without perturbation and c_2 to represent the concept weight after the perturbation. Then CFS is defined as $\text{CFS} = \|c_2 - c_1\|/\|c_1\|$. CPCS measures the cosine similarity between two concept weight vectors, which is defined as $\text{CPCS} = c_1 \cdot c_2/\|c_1\|\|c_2\|$. The smaller the value of CFS, the less the conceptual weights change after being perturbed, and the more stable the model interpretability is. The closer the value of CPCS is to 1, the higher the similarity of conceptual weights before and after perturbation and the more stable interpretability of the model. More experimental details are in the Appendix G.

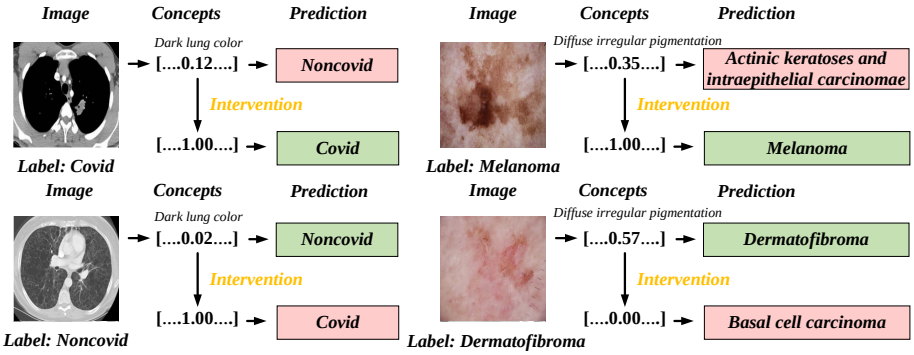

Fig. 4: Concept-intervention examples.

4.2 Utility Evaluation

Table 1 presents the accuracy results of our proposed SVCT method and the baseline approach on four datasets with different levels of perturbations. The table clearly shows that our method maintains a consistently high accuracy across all datasets without any noticeable variation or loss. This highlights the robustness of our approach in terms of accuracy preservation. Compared to Label-free CBM, our model can maintain higher accuracy while guaranteeing interpretability. Overall, the results in Table 1 show that our SVCT model successfully com-

bines high accuracy and interpretability and maintains stability over multiple datasets.

Table 3: Results on sensitivity and specificity for the baselines and SVCT w/w.o perturbation.

Method	HAM10000		Covid19-CT		BloodMNIST		OCT2017	
	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity	sensitivity	specificity
Label-free CBM	0.8878	0.9827	0.7984	**0.8608**	0.9407	0.9956	0.9750	0.9960
SVCT	**0.9899**	**0.9999**	**0.8191**	0.8037	**0.9667**	**0.9958**	**0.9950**	**0.9994**
$\rho_u = 10/255$ - LF CBM	0.6779	0.9615	0.5794	**0.9810**	0.5880	**0.9998**	0.8380	0.9880
$\rho_u = 10/255$ - **SVCT**	**0.9180**	**0.9932**	**0.7136**	0.9303	**0.8681**	0.9948	**0.9790**	**0.9923**

4.3 Stability Evaluation

Table 2 illustrates the experimental result for CFS and CPCS, assessing the stability of CBMs across various disturbance radii and comparing it with the baseline models. SVCT demonstrates superior stability concerning conceptual weights, showcasing minimal disparities pre and post-disturbance, signifying notable similarity. The prowess of SVCT in both CFS and CPCS exceeds that of the baseline model. These outcomes imply that SVCT maintains interpretability with robust resistance to perturbation, establishing it as a model with faithful explanations.

In order to represent the experimental results more intuitively, we first visualized the conceptual weight changes before and after the perturbation of each data. The results of these visualizations provide an intuitive explanation of the validity and stability of the SVCT's performance under the perturbation. The results in both Table 2 and Figure 3 amply demonstrate that, compared with the baseline model, the SVCT is a model with superior stability while keeping interpretability to perturbation. These advantages make SVCT valuable in the medical field. Secondly, we also conducted repeated experiments in several conceptual spaces to verify the validity of SVCT. Details can be found in Appendix K.

4.4 Interpretability Evaluation

Faithfulness and stability. SVCT introduces a DDS module while ensuring interpretability, which enables SVCT to provide faithful interpretations, and the results in Table 2 and Figure 3 have shown that the stability performance of SVCT performs even better under input perturbations. Experimental results indicate that SVCT is a faithful model.

Test-time intervention. We envision that in practical applications, medical experts interacting with the model can intervene to "correct" concept values that the model predicts incorrectly. During the inference process, we initially predict

Table 4: Ablation study of SVCT on DDS module. We assess the efficacy of denoising and smoothing under input perturbations.

Method	Setting		HAM10000		Covid19-CT		BloodMNIST		OCT2017	
	Denosing	Smoothing	CFS	CPCS	CFS	CPCS	CFS	CPCS	CFS	CPCS
$\rho_u = 6/255$			0.3361	0.9394	0.6761	0.7650	0.5432	0.8436	0.3625	0.9314
		✓	0.3342	0.9405	0.6490	0.7789	0.5412	0.8462	0.3516	0.9362
	✓		0.2689	0.9607	0.5698	0.8221	0.3612	0.9288	0.3367	0.9425
	✓	✓	**0.1354**	**0.9900**	**0.5555**	**0.8359**	**0.3589**	**0.9320**	**0.3257**	**0.9468**
$\rho_u = 8/255$			0.4109	0.9098	0.8114	0.6743	0.7162	0.7328	0.3812	0.9240
		✓	0.3716	0.9255	0.7258	0.7288	0.6349	0.7862	0.3724	0.9279
	✓		0.3020	0.9503	0.6556	0.7710	0.4560	0.8724	0.3574	0.9343
	✓	✓	**0.1555**	**0.9867**	**0.6446**	**0.7818**	**0.4383**	**0.8977**	**0.3459**	**0.9387**
$\rho_u = 10/255$			0.4637	0.8844	0.8943	0.6155	0.8057	0.6670	0.3949	0.9179
		✓	0.4022	0.9119	0.7856	0.6884	0.6940	0.7453	0.3869	0.9217
	✓		0.3306	0.9402	0.7157	0.7320	**0.4988**	0.8421	0.3711	0.9283
	✓	✓	**0.1725**	**0.9836**	**0.7096**	**0.7389**	0.5058	**0.8625**	**0.3620**	**0.9321**

concepts and obtain corresponding concept scores. Subsequently, we intervene by altering concept values and generating output results based on the intervened concepts. In Figure 4, we present several examples of interventions. In the example, we observed a significant darkening of the lung color, and the model gave an incorrect prediction, which, after our corrections, ended up being correct. When the model predicts correctly, we make the wrong corrections, which likewise causes the model to predict incorrectly. SVCT gives explanations that humans can understand and that humans can modify to achieve co-diagnosis. Besides, our SVCT can also improve its faithfulness in the test-time intervention under perturbations.

Sensitivity and specificity. We also conducted sensitivity and specificity experiments on four datasets. Results are shown in Table 3. Sensitivity measures the proportion of actual positive cases that are correctly identified by the model and specificity measures the proportion of actual negative cases that are correctly identified by the model. Results show that SVCT consistently outperforms the LF CBM. For the Covid19-CT dataset, while LF CBM has the highest specificity (0.8608), SVCT demonstrates a higher sensitivity (0.8191), suggesting better detection of positive cases. When perturbation ($\rho_u = 10/255$), SVCT continues to show robust performance. For example, on the HAM10000 dataset, SVCT maintains high sensitivity (0.9180) and specificity (0.9932). These results demonstrate that SVCT not only performs well under standard conditions but also maintains high accuracy and robustness in the presence of data perturbations, making it a promising method for medical image analysis.

4.5 Ablation Study

Results are shown in Table 4 and 5. The denoising diffusion model and randomized smoothing play an important role in SVCT. When we remove the denoising

Table 5: Ablation study of SVCT on DDS module. We assess the efficacy of denoising and smoothing under input perturbations.

Method	Setting		HAM10000	Covid19-CT	BloodMNIST	OCT2017
	Denosing	Smoothing				
$\rho_u = 0$			99.00%	81.23%	96.81%	99.40%
		✓	98.33%	80.54%	95.88%	99.20%
	✓		98.88%	81.09%	96.33%	99.50%
	✓	✓	**99.05%**	**81.37%**	**96.96%**	**99.50%**
$\rho_u = 10/255$			92.56%	68.22%	80.59%	95.40%
		✓	92.66%	69.10%	81.14%	97.00%
	✓		96.11%	70.03%	90.21%	98.10%
	✓	✓	**97.24%**	**71.65%**	**92.65%**	**98.48%**

diffusion model, the performance of the model suffers significantly. While removing the randomized smoothing, the model performance degradation is small. When both modules are removed at the same time, the overall performance of the model decreases more significantly compared to removing a single module. This suggests that these two modules play a key role in maintaining conceptual stability while being able to provide faithful explanations. The ablation results show that without any one of the two modules, the performance of disease diagnosis may suffer. More ablation studies about the effect of feature fusion and DDS are shown in Appendix H, indicating that each module in our SVCT is necessary and efficient. The computational cost is shown in Appendix I, implying the efficiency of our SVCT.

5 Conclusion

In this paper, we propose the Vision Concept Transformer (VCT), and further propose the Stable Vision Concept Transformer (SVCT) framework. In SVCT, we utilize ViT as a backbone, generate the concept layer, and fuse the concept features and image features. SVCT mitigates the information leakage problem caused by CBM and maintains accuracy. Comprehensive experiments show that SVCT can provide stable interpretations despite perturbations to the inputs, with less performance degradation than CBMs and maintaining higher accuracy, indicating SVCT is a more faithful explanation tool.

Acknowledgements

This work is supported in part by the funding BAS/1/1689-01-01, URF/1/4663-01-01, REI/1/5232-01-01, REI/1/5332-01-01, and URF/1/5508-01-01 from KAUST, and funding from KAUST - Center of Excellence for Generative AI, under award number 5940.

References

1. Apostolidis, K.D., Papakostas, G.A.: A survey on adversarial deep learning robustness in medical image analysis. Electronics **10**(17), 2132 (2021)
2. Chauhan, K., Tiwari, R., Freyberg, J., Shenoy, P., Dvijotham, K.: Interactive concept bottleneck models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37(5), pp. 5948–5955 (2023)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
4. Ghorbani, A., Abid, A., Zou, J.: Interpretation of neural networks is fragile (2018)
5. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020)
6. Hu, L., Liu, Y., Liu, N., Huai, M., Sun, L., Wang, D.: Seat: Stable and explainable attention (2022)
7. Ismail, A.A., Adebayo, J., Bravo, H.C., Ra, S., Cho, K.: Concept bottleneck generative models. In: The Twelfth International Conference on Learning Representations (2023)
8. Jacovi, A., Goldberg, Y.: Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? (2020)
9. Kermany, D.S., Kermany, D.S., Goldbaum, M.H., Cai, W., Valentim, C.C.S., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M.K., Pei, J., Pei, J., Ting, M.Y.L., Zhu, J., Li, C.M., Hewett, S., Hewett, S., Dong, J., Ziyar, I., Shi, A., Zhang, R., Zheng, L., Hou, R., Shi, W., Fu, X., Fu, X., Duan, Y., Huu, V.A.N., Huu, V.A.N., Wen, C., Zhang, E., Zhang, E., Zhang, C.L., Zhang, C.L., Li, O., Li, O., Wang, X., Singer, M.A., Sun, X., Xu, J., Tafreshi, A.R., Lewis, M.A., Xia, H., Zhang, K.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**, 1122–1131.e9 (2018), https://api.semanticscholar.org/CorpusID:3516426
10. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions (2020)
11. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)
12. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions (2017)
13. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
14. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models. In: The Eleventh International Conference on Learning Representations (2022)
15. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models (2023)
16. Oikarinen, T., Weng, T.W.: Clip-dissect: Automatic description of neuron representations in deep vision networks (2023)
17. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead (2019)
18. Sarkar, A., Vijaykeerthy, D., Sarkar, A., Balasubramanian, V.N.: A framework for learning ante-hoc explainable models via concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10286–10295 (2022)

19. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**(1) (Aug 2018). https://doi.org/10.1038/sdata.2018.161, http://dx.doi.org/10.1038/sdata.2018.161
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
21. Yan, A., Wang, Y., Zhong, Y., He, Z., Karypis, P., Wang, Z., Dong, C., Gentili, A., Hsu, C.N., Shang, J., et al.: Robust and interpretable medical image classifiers via concept bottleneck models. arXiv preprint arXiv:2310.03182 (2023)
22. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1) (Jan 2023). https://doi.org/10.1038/s41597-022-01721-8, http://dx.doi.org/10.1038/s41597-022-01721-8
23. Yeh, C.K., Hsieh, C.Y., Suggala, A.S., Inouye, D.I., Ravikumar, P.: On the (in)fidelity and sensitivity for explanations (2019)
24. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models (2023)
25. Zhao, J., Zhang, Y., He, X., Xie, P.: Covid-ct-dataset: a ct scan dataset about covid-19. arXiv preprint arXiv:2003.13865 (2020)
26. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2921–2929 (2016). https://doi.org/10.1109/CVPR.2016.319