Algorithm 1: utility_selection($\mathcal{X}$, $\mathcal{T}$, k, c, γ)

Input: $\mathcal{X}$ - list of image embeddings
$\mathcal{T}$ - list of text embeddings
k - number of concepts to be selected for each class, $k = \frac{K}{n}$
c - target class
$\gamma \in [0, 1]$ - threshold for Pearson's r

Output: $\mathcal{O}$ - selected text embeddings

$\mathcal{O} \leftarrow \{\}$ `// empty set`
while $|\mathcal{O}| < k$ **do**
  $\mathbf{t} \leftarrow \text{argmax}_{\mathbf{t} \in \mathcal{T}} U(\mathbf{t}, c)$
  $\mathcal{O} \leftarrow \mathcal{O} \cup \{\mathbf{t}\}$
  $\mathcal{T} \leftarrow \mathcal{T} \setminus \{\mathbf{t}\}$
  $\mathcal{R} \leftarrow \{\mathbf{t}' | \ \text{abs}\big(\rho(\mathbf{t}, \mathbf{t}')\big) > \gamma, \mathbf{t}' \in \mathcal{T}\}$
  `//` $\rho(.,.)$ `denotes Pearson's` r
  if $k - |\mathcal{O}| <= |\mathcal{T} \setminus \mathcal{R}|$ **then**
    $\mathcal{T} \leftarrow T \setminus \mathcal{R}$
  else
    $\mathcal{A} \leftarrow$ utility_selection$(\mathcal{X}, \mathcal{T}, k - |\mathcal{O}|, c, \gamma + 0.1)$ $\mathcal{O} \leftarrow \mathcal{O} \cup \mathcal{A}$
return $\mathcal{O}$

Table 1. (Left) An example of concept selection outcome by using Algorithm 1 on the HAM dataset with $k = 50$ selected concepts from a total of GPT-4 generated 760 concepts. (Right) Comparison of concept selection method for HAM dataset using our concept generation method.

Avg. word len.	Color	Shape	Size	Texture	Total
4.4	22	36	21	36	115
6.0	23	26	18	38	105
8.7	30	33	28	39	130
Total	75	95	67	113	350

Concept Selection →	Concept Utility (ours)			Submodular [23]			Label-free
CBM ↓ \ k →	10	20	50	10	20	50	CBM [14]
LaBo	73.8	75.3	76.8	73.0	73.6	74.7	72.6
AdaCBM	82.8	82.8	81.9	82.9	82.9	82.1	81.8

Fig. 1. Top-5 semantically similar concept pairs for a "Dermatofibroma" case. We show the cosine similarity between each pair using the CLIP text encoder-produced embeddings. Even if semantically similar, each pair's score is as high as we expect. However, our AdaCBM is robust to concept generation, which can achieve high performance on both concept types as shown in Table 2-(1) in the main manuscript.

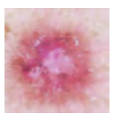
Example Image	GPT Generated Concepts	Cos. Sim.	Doctor-labeled Concepts
	1. rarely, it might show up as a dome-shaped bump on the skin	0.81	1. usually dome-shaped
	2. may appear as oval shape	0.71	2. generally round but can be oval
	3. appear as a light pink hue	0.69	3. occasionally can be pink or red
	4. may exhibit central hardening	0.55	4. can appear indurated or hardened
	5. may reach up to 10mm in diameter	0.49	5. size usually ranges from 3 to 10 mm

Table 2. Ablation study of the proposed AdaCBM model on (1) the importance of the geometrically represented quantities in terms of contribution to accuracy; (2) GPT-3/-4 generated concepts; (3) AdaCBM trained with different backbones. All results are generated on the HAM dataset. The Baseline, GPT-4, and ViT-L/14 columns are identical as they are named to the different aspects of the same baseline AdaCBM model in Table 1 in the main manuscript.

k	(1) Importance of the Geometrically Represented Quantities					(2) LLM		(3) Backbones				
								CLIP			BioMedCLIP	PLIP
	Baseline	$\|\mathbf{x}\| = 1$	$\|\mathbf{t}\| = 1$	$\|\mathbf{x}\|\|\mathbf{t}\| = 1$	$\hat{\mathbf{x}} \cdot \hat{\mathbf{t}} = 1$	GPT-3	GPT-4	ViT-L/14	ViT-B/32	ResNet-50	[25]	[8]
10	82.8	**82.9**	82.8	82.8	66.8	**82.9**	82.8	**82.8**	79.1	77.4	67.8	82.5
20	82.8	82.8	**83.2**	82.6	3.3	**82.8**	**82.8**	82.8	79.2	78.8	70.7	**82.9**
50	81.9	**82.6**	78.8	78.1	1.2	**82.4**	81.9	**81.9**	79.3	81.3	71.6	81.8