A. Experiments

A.1. Additional results

We present comprehensive results that expand on those reported in the paper. Initially, we list the hyperparameter values used in our model configuration and present the test accuracy results for varying numbers of concepts: 8, 16, 32, 64, 128, 256, one per class (1-*pc*), two per class (2-*pc*), and the entire descriptor pool. We detail the hyperparameters used and their selection in Section A.2 and report their results in Table 7.

Additionally, Table 6 includes the complete results of the three runs that form the basis for the mean and standard error reported for each dataset in Table 1.

A.2. Implementation details

Hyperparameters In Table 7, we provide an overview of the hyperparameters that configure our model for each dataset, along with their values and the empirical results. During the sampling procedure from the joint image–descriptor distribution, the transformation is calculated as formulated in Equation 2. For each dataset, we search for suitable values of $\epsilon \in \{1, 0.1, 0.01\}$ and $t \in \{1, 3, 5, 7, 10\}$. To balance the two terms in the loss function, we determine the optimal λ value from $\lambda \in \{1, 0.1, 0.01, 0.001\}$.

We also fine-tune the batch size, learning rate, random seed values, and the number of epochs for training the model during the embedding approximation learning phase ($epochs_1$), as detailed in Section 3.2, and during the training of the linear layer ($epochs_2$), as described in Section 3.4.

We obtain our score model by minimizing the objective presented in Equation 1. Training the score model involves a network with three linear layers, each having hidden dimensions of 1024. For all datasets, training is performed on the image and descriptor embeddings for 1000 epochs using the Adam optimizer with a fixed learning rate of $1e-4$. The batch size for the images remains the same as before, while the batch size for the descriptors is set to 32.

Descriptors pool filtering During the concept selection phase described in Section 3.3, we construct Sim by employing Equation 5 and retain only the top m-most similar concepts in Sim for each learnable concept. Initially, we set m to 5 and define $TopDes = \bigcup_{i=1}^{k}\{sort(Sim_i)^{(1)}, \dots, sort(Sim_i)^{(m)}\}$, where we sort Sim_i and select the top m-most similar embeddings. If the resulting pool size is greater than k, we proceed with concept selection; otherwise, we iteratively find $TopDes$ for $m_{i+1} = 2m_i$ until this condition is met. Generally, a low value of m indicates diverse learned embeddings. The reader is reffered to Table 7 for the obtained values.

Citations and rights We have thoroughly cited all datasets and research papers used in our experiments throughout our paper. The CLIP model [39] is available under the MIT license.

B. Descriptor visualizations

To gain insights into the structure of the textual descriptions, we visualize the descriptor pool along with the selected concepts that form our bottleneck. This visualization allows us to understand the diversity in the selection of the concepts.

By lowering the dimension of each embedding, we use t-SNE [56] to visualize both the embeddings of the descriptor pool and the embeddings of the selected concepts. The visualizations for the CIFAR-10, CIFAR-100, Flower, CUB, and Food datasets are presented in Figures 5 to 9. In these visualizations, each green point represents a concept from the descriptor pool, while each blue point represents a concept in the CLEAR bottleneck.

These visualizations illustrate how well the selected concepts represent the broader pool, which is vital for ensuring the robustness and generalizability of our approach. They demonstrate that our method's concept selections effectively distinguish between different conceptual areas and provide a diverse set of concepts.

Dataset	Accuracy when varying no. of concepts		
CIFAR-10	**8**	**10**	**20**
	81.25	85.02	88.14
	80.11	85.68	88.73
	82.16	81.87	90.61
CIFAR-100	**64**	**100**	**200**
	73.6	76.08	77.32
	73.71	76.12	77.31
	73.94	76.01	77.33
Flower	**32**	**102**	**204**
	87.25	90.39	90.98
	86.56	90.39	91.17
	87.54	89.80	91.17
CUB	**32**	**200**	**400**
	65.53	70.05	70.19
	65.08	70.48	69.71
	65.67	70.02	69.93
Food	**64**	**101**	**202**
	79.91	81.86	83.01
	79.83	81.64	82.40
	79.64	81.33	82.92

Table 6. Complete results of the three runs for each dataset

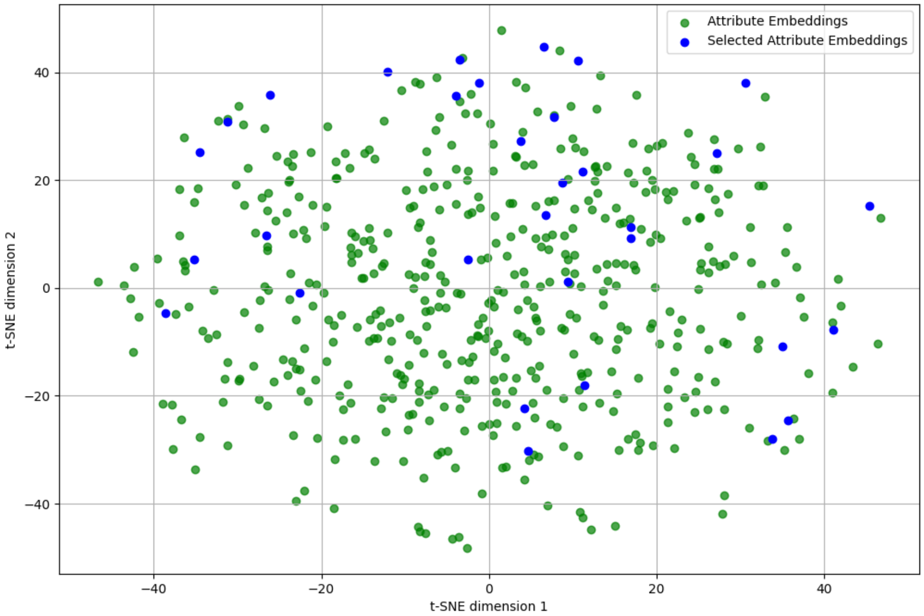

Figure 5. t-SNE visualization of CIFAR-10 descriptors

Dataset	CIFAR-10								
ϵ	1								
t	7								
λ	0.01								
batch size	4096								
learning rate	0.01								
seed	4								
$epochs_1$	1000								
$epochs_2$	2000								
no. of concepts	8	16	32	64	128	256	1-*pc*	2-*pc*	full
m	5	5	5	5	5	10	5	5	-
accuracy	81.25	87.82	92.13	93.61	94.15	94.29	85.02	88.14	94.23
Dataset	CIFAR-100								
ϵ	0.1								
t	5								
λ	0.1								
batch size	4096								
learning rate	0.01								
seed	0								
$epochs_1$	1000								
$epochs_2$	4000								
no. of concepts	8	16	32	64	128	256	1-*pc*	2-*pc*	full
m	5	5	5	5	5	5	5	5	-
accuracy	33.30	51.13	65.7	73.6	76.51	77.29	76.08	77.32	77.79
Dataset	Flower								
ϵ	0.1								
t	5								
λ	0.01								
batch size	4096								
learning rate	0.001								
seed	1								
$epochs_1$	2000								
$epochs_2$	20000								
no. of concepts	8	16	32	64	128	256	1-*pc*	2-*pc*	full
m	5	5	5	5	5	5	5	5	-
accuracy	59.60	80.19	87.25	89.51	90.29	91.17	90.39	90.98	91.37
Dataset	CUB								
ϵ	1								
t	10								
λ	1								
batch size	32								
learning rate	0.01								
seed	0								
$epochs_1$	5000								
$epochs_2$	8000								
no. of concepts	8	16	32	64	128	256	1-*pc*	2-*pc*	full
m	5	5	5	5	5	5	5	5	-
accuracy	32.01	51.81	65.53	69.96	70.29	69.95	70.05	70.19	66.98
Dataset	Food								
ϵ	1								
t	1								
λ	1								
batch size	4096								
learning rate	0.01								
seed	0								
$epochs_1$	200								
$epochs_2$	4000								
no. of concepts	8	16	32	64	128	256	1-*pc*	2-*pc*	full
m	5	5	5	5	5	5	5	5	-
accuracy	39.10	58.58	74.42	79.91	81.61	82.59	81.86	83.01	82.55

Table 7. Hyperparameter values and full results on varying numbers of concepts.

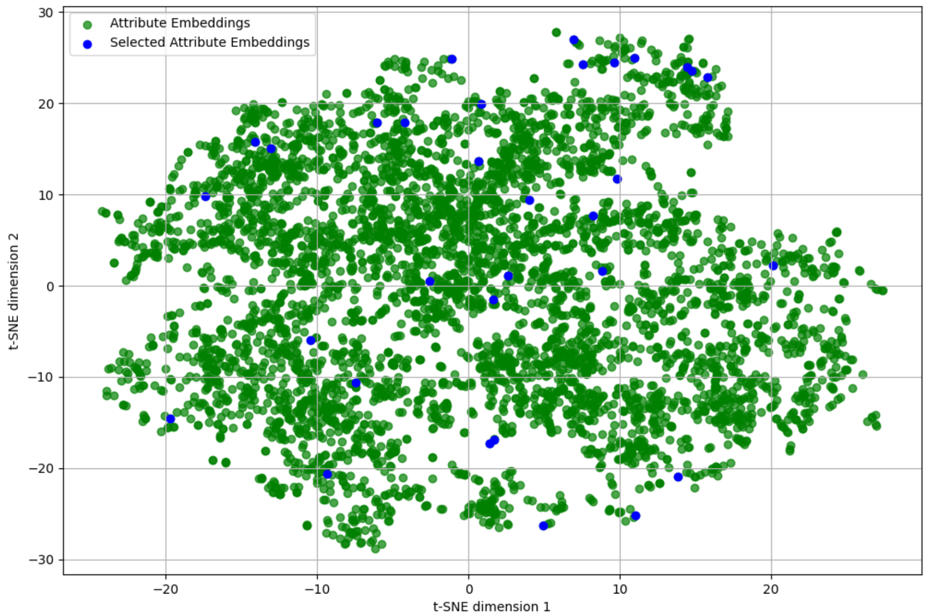

Figure 6. t-SNE visualization of CIFAR-100 descriptors

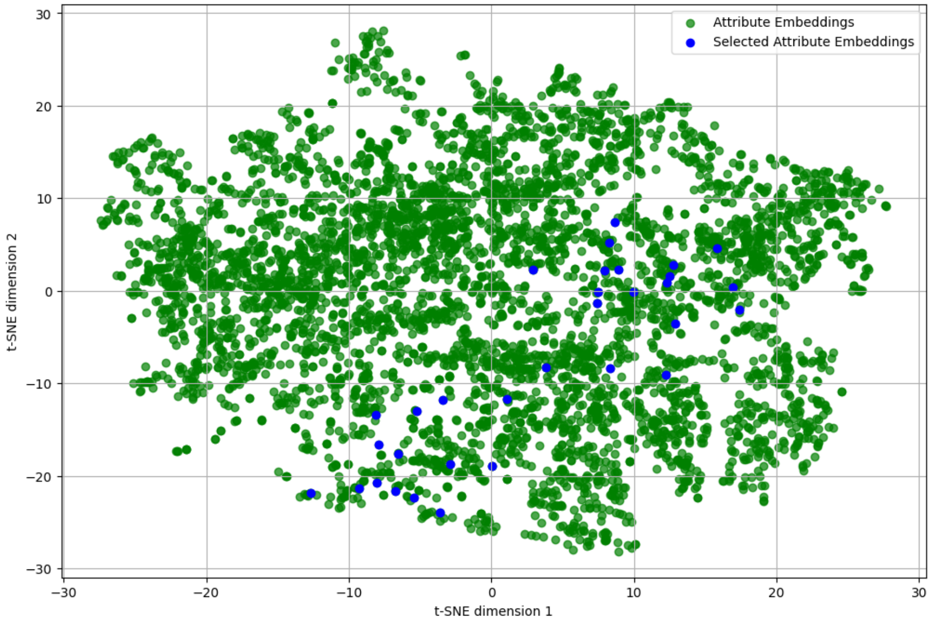

Figure 7. t-SNE visualization of Flower descriptors

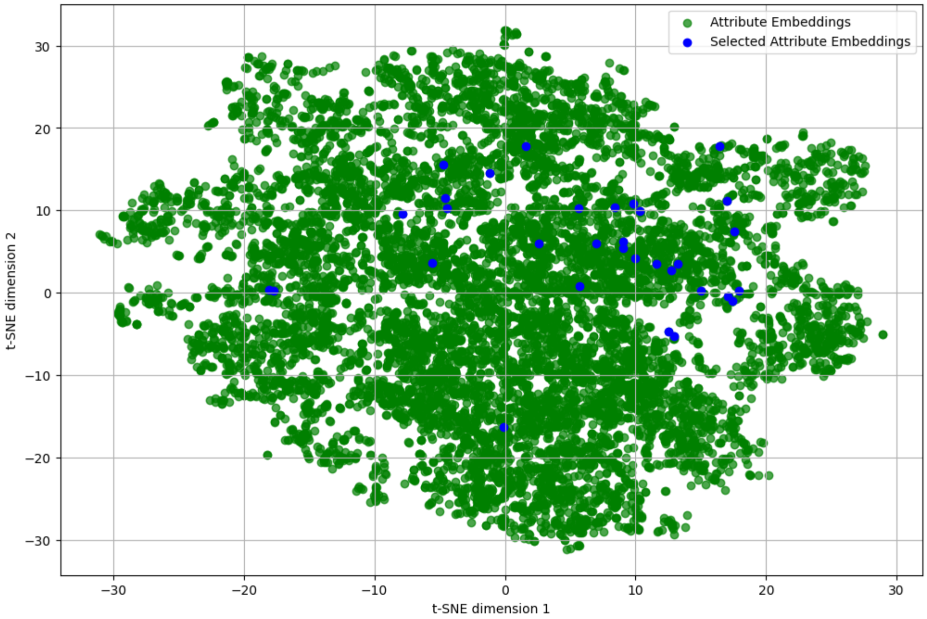

Figure 8. t-SNE visualization of CUB descriptors

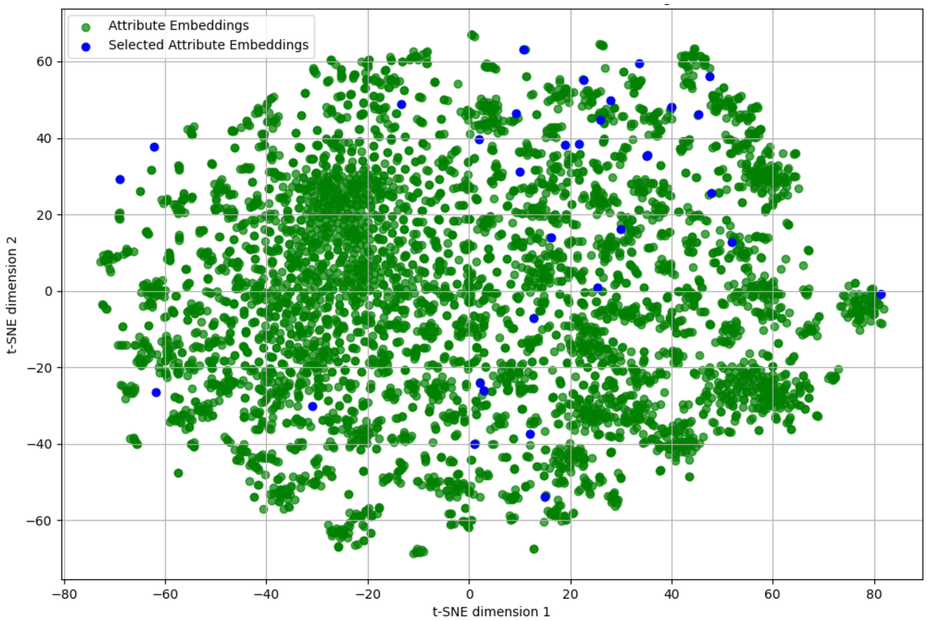

Figure 9. t-SNE visualization of Food descriptors