Rethinking Concept Bottleneck Models: From Pitfalls to Solutions

Supplementary Material

A. Supplementary

A.1. ImageNet100 Class List

Table 4. List of 100 ImageNet classes used in ImageNet100.

n02869837	n01749939	n02488291	n02107142
n13037406	n02091831	n04517823	n04589890
n03062245	n01773797	n01735189	n07831146
n07753275	n03085013	n04485082	n02105505
n01983481	n02788148	n03530642	n04435653
n02086910	n02859443	n13040303	n03594734
n02085620	n02099849	n01558993	n04493381
n02109047	n04111531	n02877765	n04429376
n02009229	n01978455	n02106550	n01820546
n01692333	n07714571	n02974003	n02114855
n03785016	n03764736	n03775546	n02087046
n07836838	n04099969	n04592741	n03891251
n02701002	n03379051	n02259212	n07715103
n03947888	n04026417	n02326432	n03637318
n01980166	n02113799	n02086240	n03903868
n02483362	n04127249	n02089973	n03017168
n02093428	n02804414	n02396427	n04418357
n02172182	n01729322	n02113978	n03787032
n02089867	n02119022	n03777754	n04238763
n02231487	n03032252	n02138441	n02104029
n03837869	n03494278	n04136333	n03794056
n03492542	n02018207	n04067472	n03930630
n03584829	n02123045	n04229816	n02100583
n03642806	n04336792	n03259280	n02116738
n02108089	n03424325	n01855672	n02090622

A.2. Additional Experimental Setup Details

In Tab. 5, we report the learning rates (η) used for optimizing the concept encoder and classifier, along with the distillation temperature and the α/β weights for balancing the classifier loss terms. The table also indicates whether a learning rate scheduler is used. When needed, we employ a cosine annealing schedule with a minimum learning rate of 0.0001 and a cycle length of 20 epochs. Hyperparameters are provided per dataset, and the same settings are used across different vision backbones and VLMs.

The Non-linear CBM in Fig. 5 and the Distilled CBM in Fig. 6 correspond to the same underlying model. For the Linear CBM and Vanilla CBM ablation experiments, we use the same hyperparameters for consistency.

A.3. Irrelevant and Random Words Concept Sets

We provide the randomly selected concepts from our irrelevant and random words concept sets in Tab. 6. These concept sets are used throughout all experiments presented in Fig. 2, Fig. 3, Fig. 5 and in Tab. 1. The Roman Law concepts were extracted from the corresponding Wikipedia page, while the random strings were sampled uniformly at random, as described in Sec. 5.

Table 5. Hyperparameter settings for each dataset, including learning rates (η) and scheduler usage for the concept encoder and classifier, along with the α and β loss weights and distillation temperature Identical settings are used across all vision backbones and VLMs.

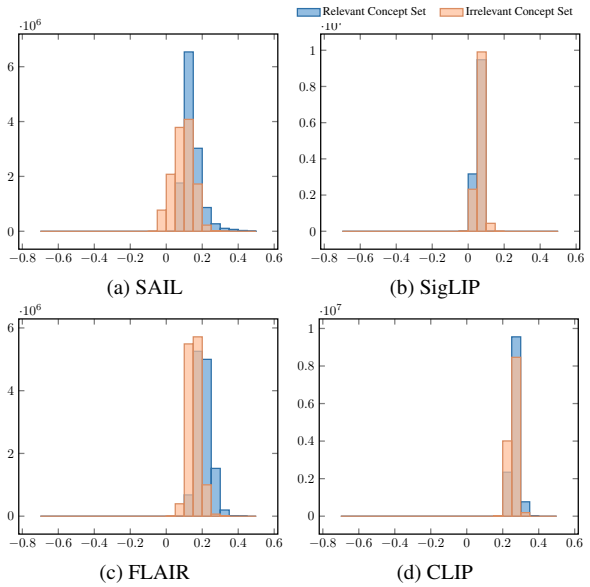
Dataset	**Concept Encoder**		**Classifier**				
	η	Scheduler	α	β	Temperature	η	Scheduler
ImageNet100	0.001	✓	1	1	2	0.0001	-
CUB200	0.0001	-	1	0.5	4	0.001	✓
Places365	0.0001	-	1	2	2	0.001	✓
CIFAR100	0.001	✓	1	1	4	0.0001	-

Table 6. Sample of 30 concepts from the two concept sets used in our experiments. Roman Law concepts extracted from the content of the relevant Wikipedia page, and a set of randomly generated strings.

Roman Law Concepts		**Random Words**	
Jurisprudential	historians	gsankm	sjklrveogjc
administrative	impede	ehxrpq	pdbhhidbnvjig
bureaucratization	individual	swsbkzguqjil	cygmao
centuries	iudicum	odbhhqr	foqmegkdcp
civil	knowledge	rwkkmbezusn	fdqumo
delegation	latitude	qkfns	hkeeyhw
descendants	laws	ntqwbpkwald	kzhqpxhr
disappear	legal	jnplmlek	sgwftaole
ecclesiastical	magistrate	qtxvhsjlex	esdyegcbw
emperors	mancipatio	qfggdblkkli	zghijqngltcoo
enjoy	military	nckusvorw	tihgu
faithful	more	znnufioys	dmcmmjvd
features	nations	kvqpasqqkgt	ghhtwblxlmphi
formularies	normal	ydatcmpeqsmnf	ulcgplindw
grants	obvious	zebwdzhjfx	cynrhdme

A.4. Concept Activation Distributions of Different VLMs

We compute concept activations on the ImageNet100 dataset using both the GPT-generated concept set as the relevant concepts and a set of random strings as the irrelevant concepts. The evaluated VLMs include CLIP, SigLIP, SAIL, and FLAIR. For each model, we report the distribution of the top-100 concept activations per image in Fig. 8. Across the four VLMs, the activation distributions differ substantially in both scale and spread. SAIL exhibits the



(a) SAIL (b) SigLIP (c) FLAIR (d) CLIP

Figure 8. Histogram of concept activations computed with different VLMs for the relevant and irrelevant concept sets on ImageNet100, highlighting differences in activation distributions across models when evaluated on meaningful versus random concepts.

Table 7. Robustness of the goodness of concepts metric to varying cutoff values on ImageNet100.

	Task Agnostic ↓		
Concept Set	**Top-10**	**Top-100**	**Top-250**
Relevant	**1.751**	**3.660**	**4.521**
Irrelevant	2.254	4.504	5.376
Random	2.284	4.565	5.465

widest distribution, with both relevant and irrelevant concepts producing high activations, whereas SigLIP and CLIP show much more compact distributions. Overall, these results highlight the differences in how VLMs encode concepts.

A.5. Goodness of Concepts with Varying Cutoff Values

We use a cutoff of 100 concepts that can be consistently applied to both ImageNet100 and CUB200, whose concept sets contain 4,751 and 370 concepts, respectively. Importantly, the goodness-of-concepts metric is largely insensitive to the specific choice of cutoff: relevant concept sets consistently exhibit lower entropy, whereas irrelevant concept sets yield higher entropy across different cutoff values. In Tab. 7, we provide a quantitative evaluation of goodness of concepts on ImageNet100 with varying cutoff values. We observe that varying cutoff values yields robust behavior.

Table 8. Task-agnostic goodness of concepts: suitability of different concept sets for different datasets. Relevant concepts are GPT-generated in ImageNet100 and Places365 and ground-truth attributes in CUB200. The irrelevant concept set consists of Roman Law terminology. Random strings denote non-word concepts. Lower entropy indicates better alignment.

ImageNet100

Concept Set	CLIP	SAIL	FLAIR	SigLIP
Relevant	4.479	3.660	4.545	4.494
Irrelevant	4.480	4.504	4.511	4.534
Random words	4.555	4.565	4.565	4.522

CUB200

Concept Set	CLIP	SAIL	FLAIR	SigLIP
Relevant	4.456	4.305	4.459	4.463
Irrelevant	4.422	4.491	4.533	4.534
Random words	4.542	4.553	4.577	4.491

Places365

Concept Set	CLIP	SAIL	FLAIR	SigLIP
Relevant	4.482	3.077	4.534	4.512
Irrelevant	4.472	4.505	4.531	4.539
Random words	4.557	4.566	4.563	4.600

A.6. Goodness of Concepts Results with Various VLMs

The goodness of concept metric is calculated across ImageNet100, CUB200 and Places365 datasets using several VLMs. Results are reported in two settings: task-agnostic in Tab. 8 and task-specific in Tab. 9.

Across all datasets and models, the metric consistently demonstrate a clear separation between relevant and irrelevant concepts. Relevant concepts achieve the lowest entropy, indicating stronger alignment between the concepts and the images in both settings.

This trend is especially pronounced for SAIL, which also exhibits a distinctive concept activation distribution in Fig. 8. SAIL yields markedly lower entropy values for relevant concepts in both ImageNet100 and Places365.

Overall, these results confirm that the goodness of concepts metric reliably captures concept set suitability across a range of VLMs and diverse domains.

A.7. Refining concept set via Goodness of Concepts

In Fig. 9, we demonstrate that the goodness of concepts metric enables concept set refinement. We create a mixed concept set with 75% CUB attributes and 25% random words. The figure above shows entropy reduction over removal iterations. The red line represents random concept

Table 9. Task-specific goodness of concepts: suitability of different concept sets for different datasets. Relevant concepts are GPT-generated in ImageNet100 and Places365 and ground-truth attributes in CUB200. The irrelevant concept set consists of Roman Law terminology. Random strings denote non-word concepts. Lower entropy indicates better alignment.

ImageNet100

Concept Set	CLIP	SAIL	FLAIR	SigLIP
Relevant	4.517	3.738	4.578	4.522
Irrelevant	4.529	4.545	4.572	4.559
Random words	4.576	4.579	4.589	4.556

CUB200

Concept Set	CLIP	SAIL	FLAIR	SigLIP
Relevant	4.456	4.383	4.459	4.463
Irrelevant	4.422	4.526	4.533	4.534
Random words	4.542	4.565	4.577	4.491

Places365

Concept Set	CLIP	SAIL	FLAIR	SigLIP
Relevant	4.549	2.889	4.579	4.512
Irrelevant	4.551	4.547	4.583	4.540
Random words	4.591	4.578	4.592	4.600

removal (averaged over 10 trials), while the blue line shows entropy-guided removal (selecting concepts whose removal minimizes entropy). Entropy-guided removal achieves consistently lower entropy, indicating more concentrated and meaningful concept activations. Crucially, entropy-guided removal preferentially eliminates random words: 17.9% of removed concepts came from the random-word subset, compared to only 13.7% for random removal. This demonstrates that goodness of concepts effectively identifies and removes irrelevant concepts, validating its utility for concept set refinement.

A.8. Ablations

To inspect the disentangled effects of non-linearity and distillation, we conduct an ablation experiment. As shown in Tab. 10, linearity is the primary factor affecting performance. However, non-linearity is required for a trustworthy CBM (*cf*. Fig. 5). To reconcile this trade-off, we improve the accuracy of non-linear models through knowledge distillation, which consistently recovers a substantial portion of the performance gap.

A.9. Extended Interpretability Results on Different Datasets

The extended version of Fig. 7 is provided in the Fig. 10 where we report the top-5 activated concepts for random

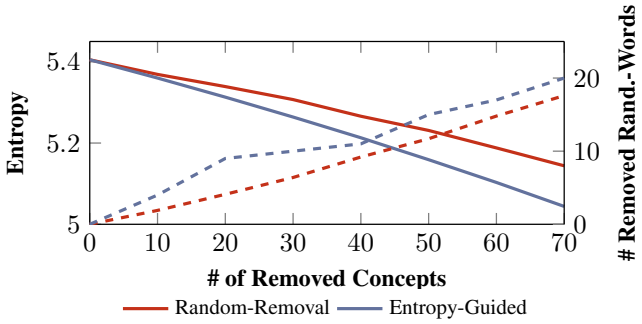


Figure 9. Concept set refinement. Entropy-guided removal (blue) outperforms random removal (red, averaged over 10 trials) and preferentially eliminates random concepts, demonstrating the effectiveness of the goodness-of-concepts metric. Solid lines correspond to the entropy (left y-axis), while dashed lines correspond to the number of removed random words (right y-axis).

Table 10. Ablation study disentangling the effects of linearity and knowledge distillation.

Configuration	Accuracy (%)
Linear	87.96
Linear + Distillation	90.37
Non-linear	82.27
Non-linear + Distillation	85.56

classes. The best performing vision backbone and VLM combinations for each dataset are as follows: Perception Encoder + SigLIP for ImageNet100, Perception Encoder + CLIP for all remaining datasets.

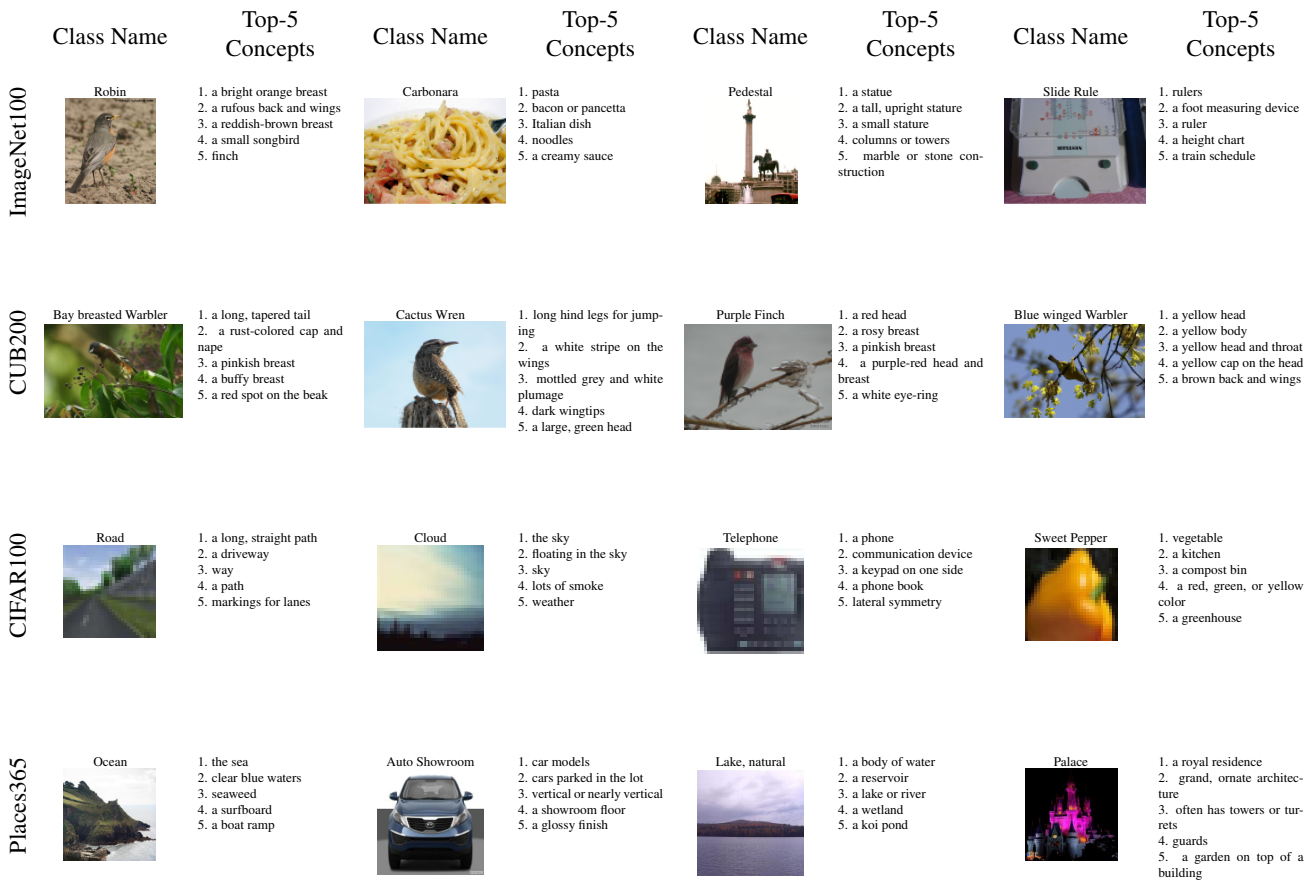


Figure 10. Illustration of the Top-5 concepts, ranked by their contribution to the final prediction for randomly selected classes, along with representative images from ImageNet100, CIFAR100, CUB200, and Places365, highlighting model interpretability across diverse domains.