Rebuttal Supplementary Material

A. Ablation Studies

Confidence threshold	CUB200		Places365	
Metrics	Acc@5	Avg. Acc.	Acc@5	Avg. Acc.
0.1	75.75%	75.75%	41.84%	42.50%
0.15	75.75%	75.73%	41.84%	42.51%
0.20	75.73%	75.73%	41.25%	42.15%

Table A.1. Accuracy at NEC@5 and Average accuracy for different confidence threshold T.

A.1. Ablation study for confidence threshold

Confidence threshold T in Eq 2 filters concepts with bounding boxes' confidence less than T. In this experiment, we study the affect of T on the VLG-CBM's accuracy. The results are shown in Table A.1. We observe that the accuracy at NEM@5 and average accuracy first increases (or stays constant) and then decreases. We attribute this effect to to the fact that as T increases, the number of false-positive decreases leading to better learning of concepts, however, as the number of annotations available for learning a concept decreases.

B. Evaluating annotations from Grounding DINO

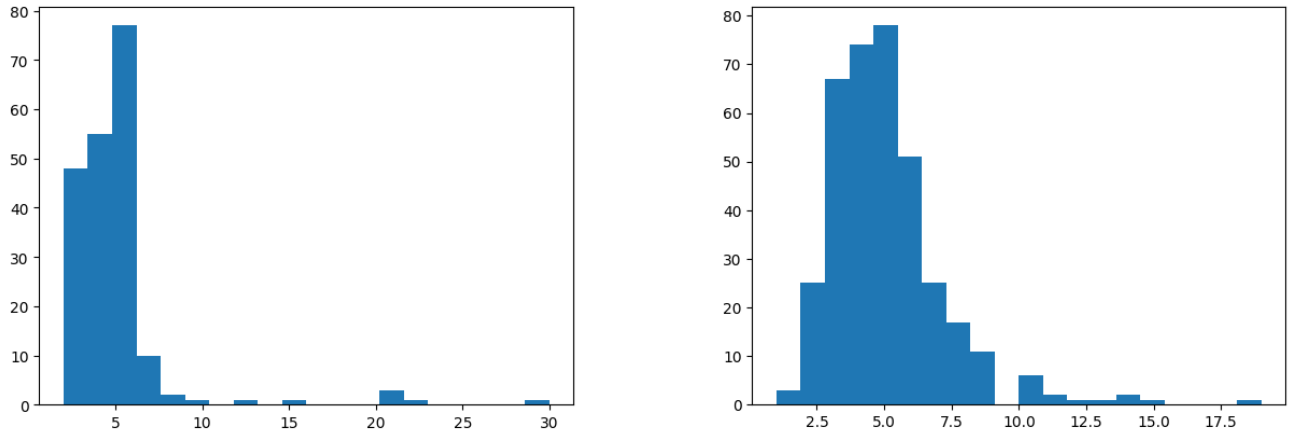
This section quantitatively evaluates concept annotations obtained from Grounding DINO. We use CUB dataset for comparison which contains ground-truth for fine-grained concepts present in each image. We use the label set from Koh et al. (2020) which has 1:1 mapping with the ground-truth concepts in the CUB dataset. We use precision and recall metric to measure the quality of annotations from Grouding DINO for each concepts. Table B.1 present mean precision and mean recall value at different confidence threshold. We observe that the obtained annotations have a very high recall i.e if the concept is present in the image, grounding DINO is able to retrieve the object. The precision is also sufficiently high though it suffers from a relatively higher false-positive detection rate compared to false-negative detection rate. However, as demonstrated in our qualitative and quantitative studies (Table 2, Fig 4, E.2, E.1) the effect of false-positive is minimal and VLG-CBMis able to faithfully represent concepts in the Concept Bottleneck Layer.

Confidence threshold	Mean Precision	Mean Recall
0.10	0.7150 ± 0.07	0.9930 ± 0.08
0.15	0.7156 ± 0.07	0.9693 ± 0.11
0.20	0.7121 ± 0.10	0.8713 ± 0.21

Table B.1. Quantitative evaluation of concepts obtained from Grounding DINO model with Mean Precision and Recall for concepts at different confidence thresholds.

C. Distribution of nonzero weights among class

The NEC metric controls the average number of non-zero weights among classes. Further, we study the distribution of non-zero weight numbers between different classes. We choose our VLG-CBM model trained on CUB and places365 datasets, which have 200 and 365 classes, respectively, and plot the distribution of non-zero weights. Both models are trained to have NEC=5 The results are shown in Figure C.1. The figure suggests most classes have non-zero weight numbers around 5, while a small number of classes utilize more concepts to make decisions.



(a) CUB

(b) Places365

Figure C.1. distribution of non-zero weight numbers from CUB and Places365 dataset. The models are trained to have NEC=5.

D. Constructing model with specified NEC

In this section, we discuss how to construct models with specified NEC. When using methods with dense final layers (e.g. (Yan et al., 2023)), controlling NEC is simply controlling total number of concepts in the concept set. Hence, below we mainly focus on models with sparse final layers.

When training the final linear layer, larger lambda(regularization strength) pushes the model to be sparser. Hence, we utilize GLM-SAGA(Wong et al., 2021), which allows us to obtain a regularization path consists of different lambdas. To be more specific, we choose a λ_{max} and train models with λ in $[\lambda_{min} = \lambda_{max}/500\lambda_{max}]$, and take 50 λ evenly from the interval in log space. Then, we choose the weight matrix with the closest NEC and pruning the weights from smallest magnitude to largest to enforce strict NEC. Hence, the actual NEC is enforced to be exactly as prespecified ones.

E. Visualizing Top concepts for dataset

This section presents extended version of Table 4 visualizing top-5 images for randomly picked concepts for CUB and Places365 dataset. The results are shown in Fig E.1 and E.3 for Places365 and Fig E.4 for CUB.

F. Additional case study examples

F.1. Negative concepts in reasoning

In LF-CBM (Oikarinen et al., 2023) and our VLG-CBM, normalization is applied on concept logits before the final decision layer. Hence, a negative value of concept logits indicates corresponding concept does not appear in the image. Following LF-CBM, we mark these concepts as "NOT {concept}" in explaining the decision. To study the frequency of this negative reasoning, we count the times these negative concepts appear in top-5 contributing concepts on CUB dataset. The results show that, for VLG-CBM, 162 out of 28950(0.56%) reasonings are through negative concepts. For comparison, LF-CBM utilizes 6687 out of 28950(23.10%) negative reasoning.

F.2. Impact of NEC

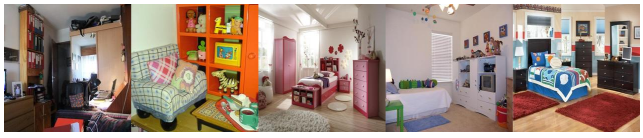
The study in Section 5.3 shows that our VLG-CBM provides more interpretable decisions than baseline methods. To better understanding where these advantages comes from, we conduct a further study to set the baselines with NEC=5 and compare the decision interpretation, see Figures F.2 to F.4The results suggest setting NEC=5 alleviate the problem from non-top-5 concepts. However, wrong/inaccurate/less useful explanations still exist.

(a) Concept: Armrest

(b) Concept: Balance beam

(c) Concept: Barren Landscape

(d) Concept: Baseball

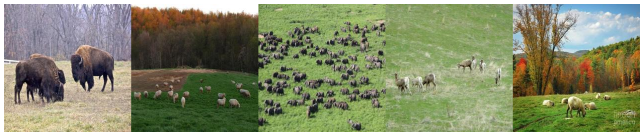


(e) Concept: Bookcase

(f) Concept: Bar or counter

(g) Concept: Statue

(h) Concept: Spoon



(i) Concept: Used for grazing animals

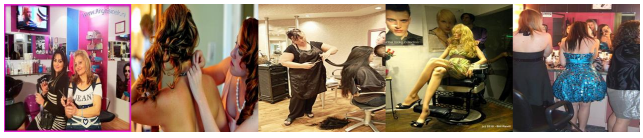
(j) Concept: Pool Table

(k) Concept: Shelves of Sewing supplies

(l) Concept: Plastic table and chairs

(m) Concept: Steering Wheel

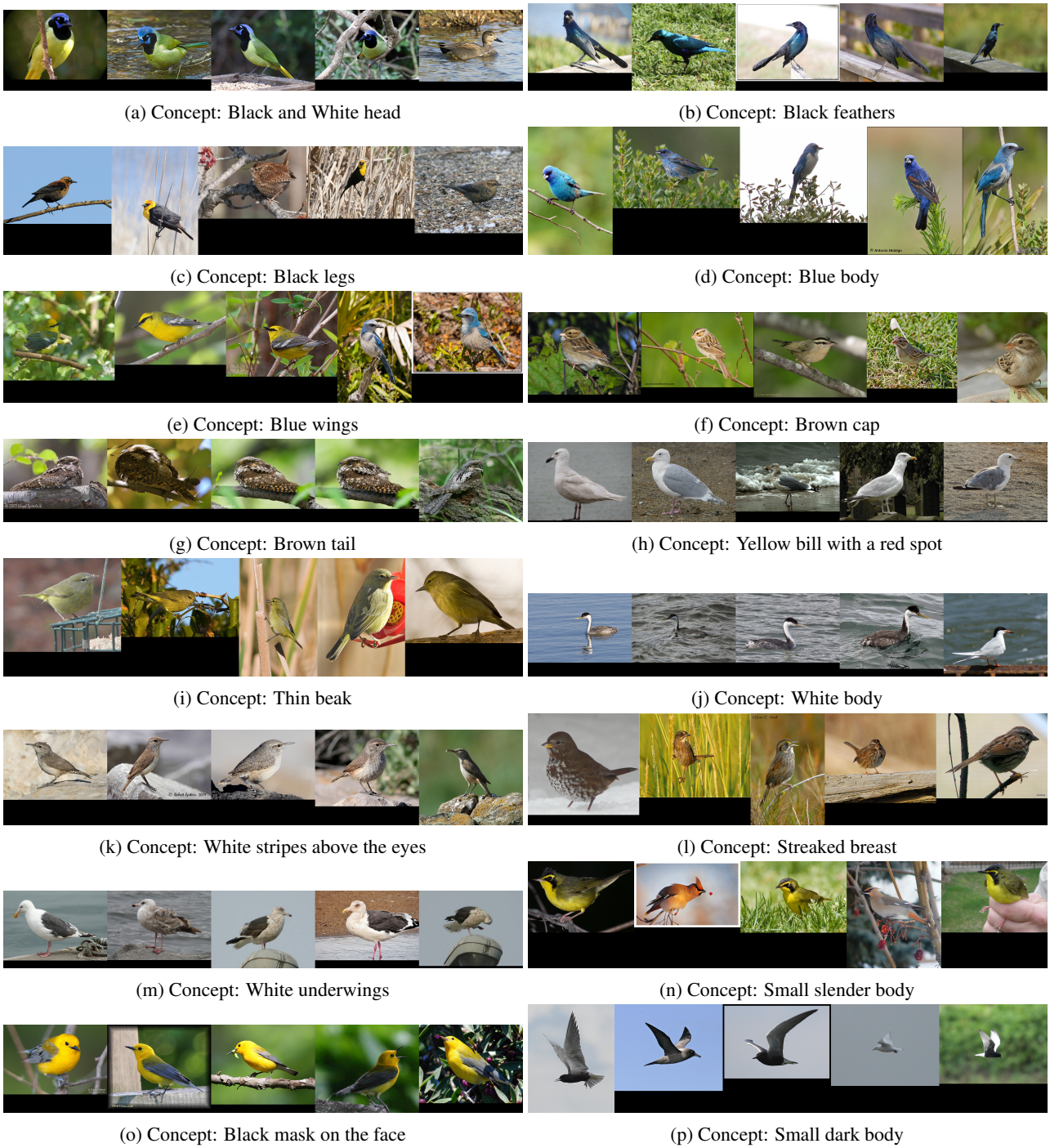
(n) Concept: Vacation



(o) Concept: Barbershop

(p) Concept: Bank

Figure E.1. Top-5 activating images for randomly selected Places365 concepts



(a) Concept: Black and White head

(b) Concept: Black feathers

(c) Concept: Black legs

(d) Concept: Blue body

(e) Concept: Blue wings

(f) Concept: Brown cap

(g) Concept: Brown tail

(h) Concept: Yellow bill with a red spot

(i) Concept: Thin beak

(j) Concept: White body

(k) Concept: White stripes above the eyes

(l) Concept: Streaked breast

(m) Concept: White underwings

(n) Concept: Small slender body

(o) Concept: Black mask on the face

(p) Concept: Small dark body

Figure E.2. Top-5 activating images for randomly selected CUB concepts

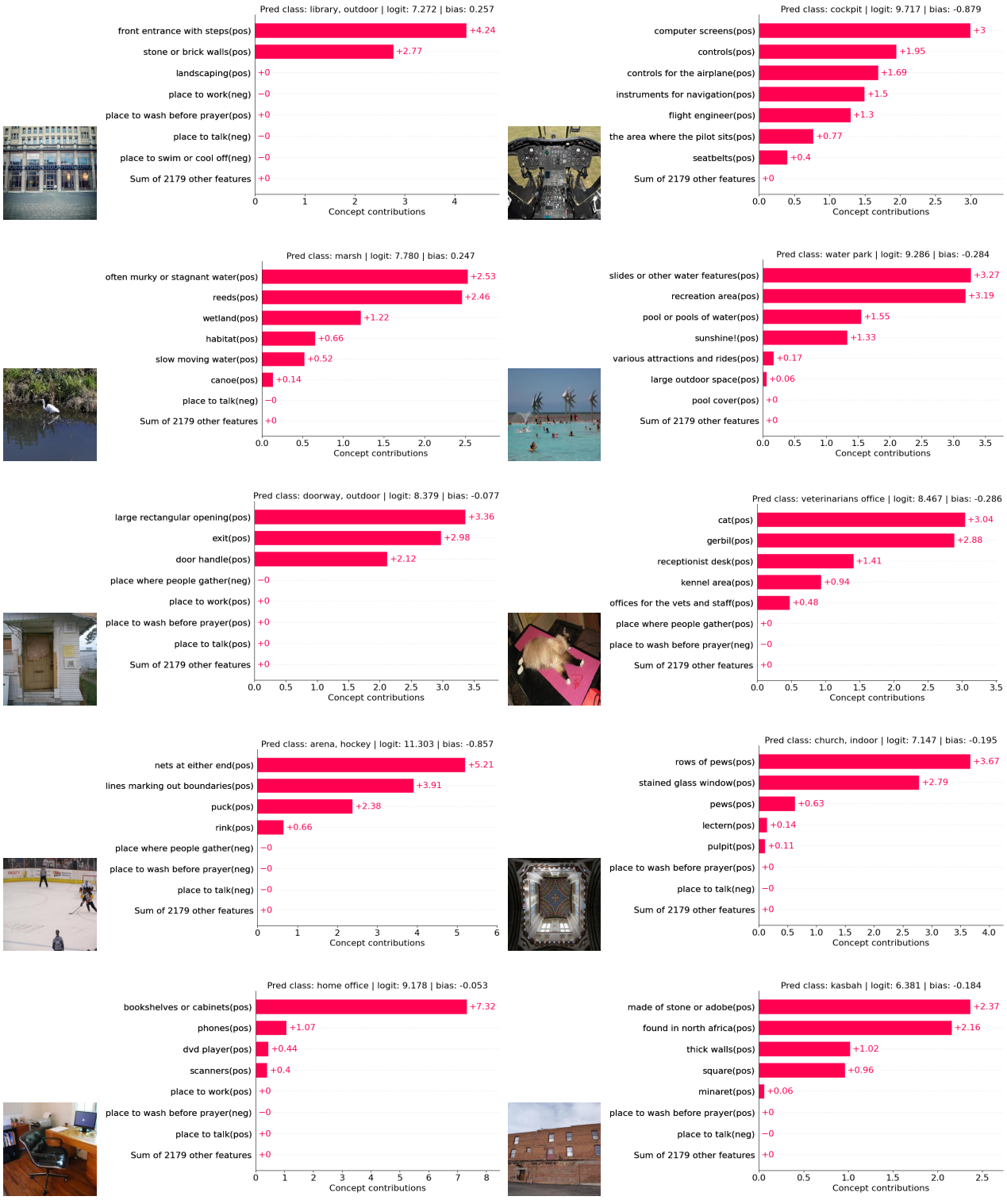


Figure E.3. Randomly selected explanations for Places365

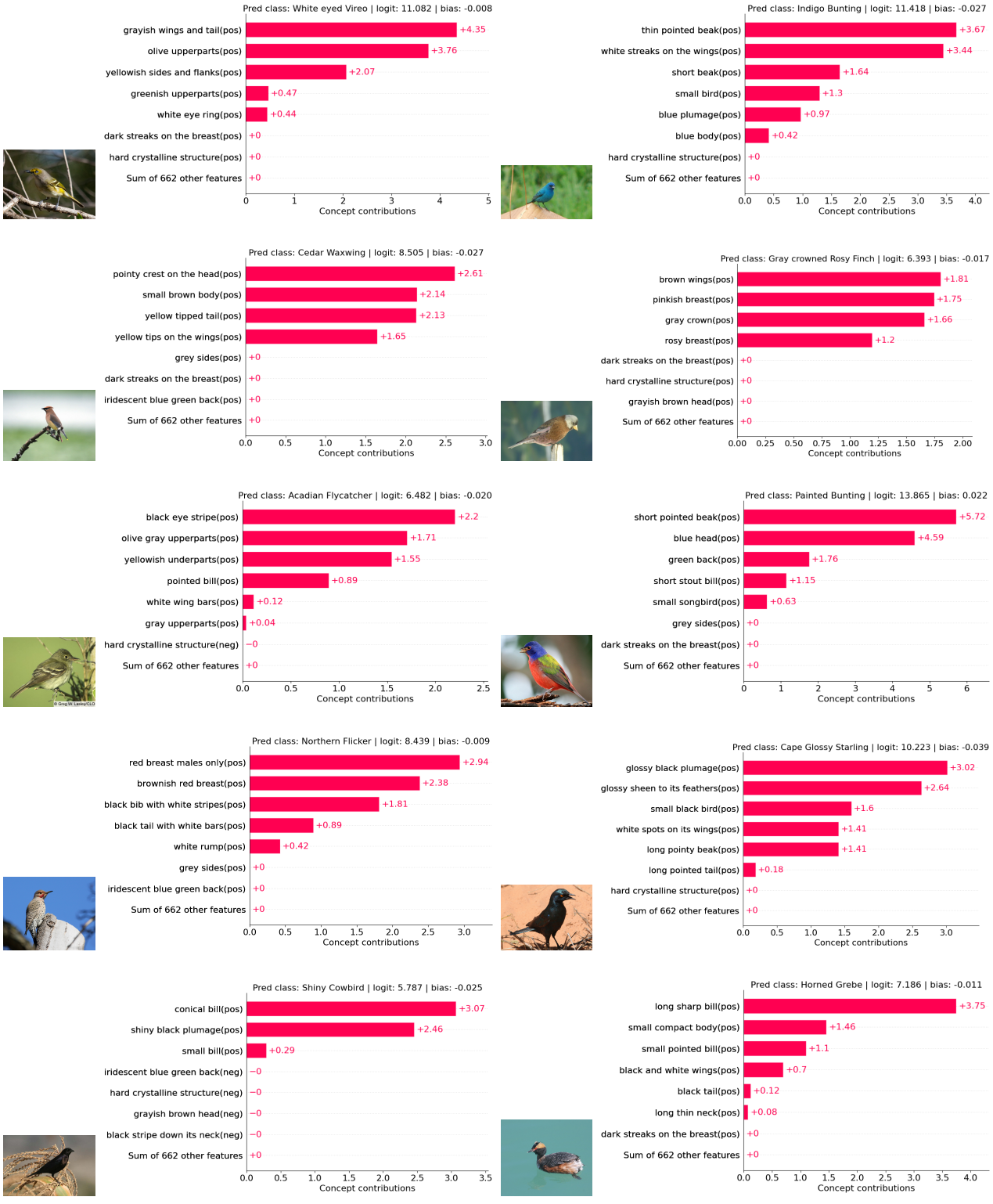


Figure E.4. Randomly selected explanations for CUB

Ground truth: **Bobolink**.
VLG-CBM prediction: **Bobolink**

1. black v on the back(4.30)
2. black and white striped head(1.94)
3. black head and back(0.27)
4. small round body(0.07)
5. NOT white stripes above the eyes(0.01)

Sum of other concepts: (0.03)

Figure F.1. Image 307: An example of negative reasoning of VLG-CBM

LF-CBM

1. NOT a brown and white color scheme (1.77)
2. NOT white and black coloration (1.66)
3. iridescent feathers (1.37)
4. NOT a black bib with white stripes (1.29)
5. NOT a black and white color scheme (1.01)

Sum of other concepts (4.22)

LF-CBM (NEC=5)

1. NOT a brown and white color scheme(2.39)
2. NOT a black bib with white stripes(1.08)
3. NOT a black and white color scheme(0.61)
4. iridescent feathers(0.27)
5. NOT yellowish-brown wings(0.25)

Sum of other concepts: (0.00)

LM4CV

1. Color - Blue head, olive back, yellow underparts (101.08)
2. grayish head, back, wings and tail with blue highlights (94.03)
3. bright blue and orange plumage (91.44)
4. large red bill with a slightly hooked tip (89.09)
5. white rump patch at the base of the tail (59.69)

Sum of other concepts (-171.50)

LM4CV (NEC=5)

1. bright reddish brown head, crown and back of neck.(382.61)
2. bright yellow, green and blue plumage(95.61)
3. bright yellow throat, breast, and flanks with black bars (51.36)
4. Broad tail that is shorter than other pelican species (-36.48)
5. Mottled brown on the nape, mantle, and scapulars(-243.90)

Sum of other concepts: (0.00)

LaBo

1. beautiful bird with a brightly colored body (0.02)
2. small, plump songbird with a short tail and a pointed bill (0.02)
3. beautiful bird with a brightly colored plumage (0.02)
4. one of the most beautiful north american songbirds (0.02)
5. colors are very vibrant and beautiful (0.02)

Sum of other concepts (41.25)

LaBo(NEC=5)

1. beautiful little bird with a very colorful plumage(3.98)
2. very colorful bird, with a lot of blue and green(0.43)
3. very pretty and very colorful(0.41)
4. NOT white stripes on white stripes on brown(0.22)
5. known as the "rainbow jay" due to its bright plumage(0.11)

Sum of other concepts: (0.00)

Figure F.2. Comparing baselines with different NECs

LF-CBM

1. NOT dark wingtips(2.07)
2. white breast with brown spots(1.93)
3. NOT a dark brown or black color(1.25)
4. yellow feet(1.12)
5. NOT long, blue-gray wings(0.81)

Sum of other concepts: (5.12)

LF-CBM (NEC=5)

1. NOT dark wingtips(1.80)
2. NOT long, blue-gray wings(0.69)
3. yellow feet(0.66)
4. a red face(0.08)
5. a Scarlet-red body(0.00)

Sum of other concepts: (0.00)

LM4CV

1. red, black and white feathers(93.11)
2. bright red head and breast(84.05)
3. bright red head and nape(75.39)
4. bright red crescent below its beak (63.14)
5. White neck with a black collar and chestnut red head and breast(60.52)

Sum of other concepts: (-172.32)

LM4CV (NEC=5)

1. bright reddish brown head, crown and back of neck.(344.38)
2. bright yellow, green and blue plumage(87.41)
3. bright yellow throat, breast, and flanks with black bars (39.26)
4. Broad tail that is shorter than other pelican species (-34.77)
5. Mottled brown on the nape, mantle, and scapulars(-227.34)

Sum of other concepts: (0.00)

LaBo

1. male goldfinch is the more brightly colored of the sexes, with(0.02)
2. seen in flocks of other goldfinches(0.02)
3. often forming flocks with other goldfinches(0.02)
4. visit bird tables and feeders(0.02)
5. young goldfinches are drabber than adults, with brownish plumage(0.02)

Sum of other concepts: (41.69)

LaBo(NEC=5)

1. closely related to the goldfinch(3.83)
2. often forming flocks with other goldfinches(1.16)
3. NOT from alaska and canada to the southwestern united states(1.10)
4. often forming flocks with other goldfinches and similar small birds(0.17)
5. NOT found in eastern and central united states year-round(0.03)

Sum of other concepts: (0.02)

Figure F.3. Comparing baselines with different NECs

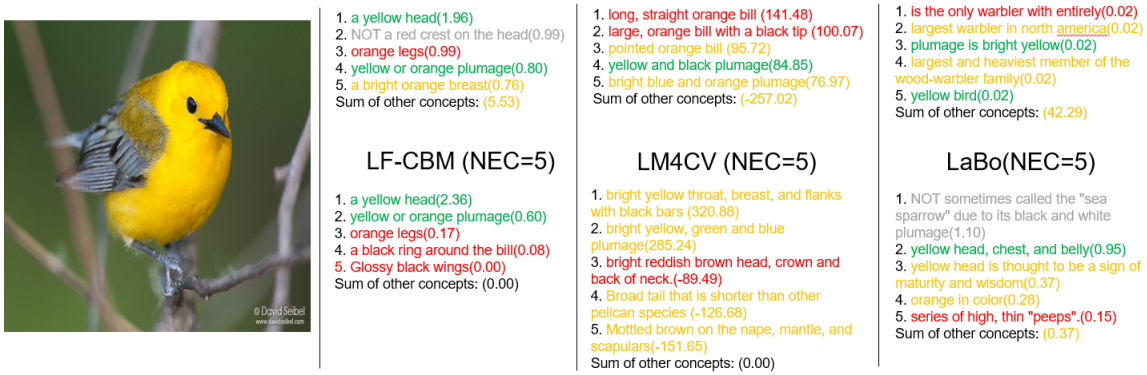


Figure F.4. Comparing baselines with different NECs