A. New Datasets

Dataset Statistics. We collect three new domain adaptation datasets, *i.e*. AwA2-clipart, LADA-Sculpture, and LADV-3D based on previous attribute-annotated datasets [1, 3] in this paper. Table. 1 are some basic dataset statistics. Fig. 1 has shown some samples from each dataset. To visualize the domain shifts in these three new datasets, we use CLIP ViT-L/14 to extract image features for images in these datasets and use the TSNE tools to visualize all image features.

Dataset Statistics	AwA2-clipart	LADA-Sculpture	LADV-3D
Data field	animals	animals	vehicles
Visual domains	photo,clipart	real,sculputre	real, 3d renders
Number of images	37328,5319	13240,2162	17080,3587
Number of categories	50	50	50

Table 1. Some dataset statistics of AwA2-clipart, LADA-Sculpture, LADV-3D.

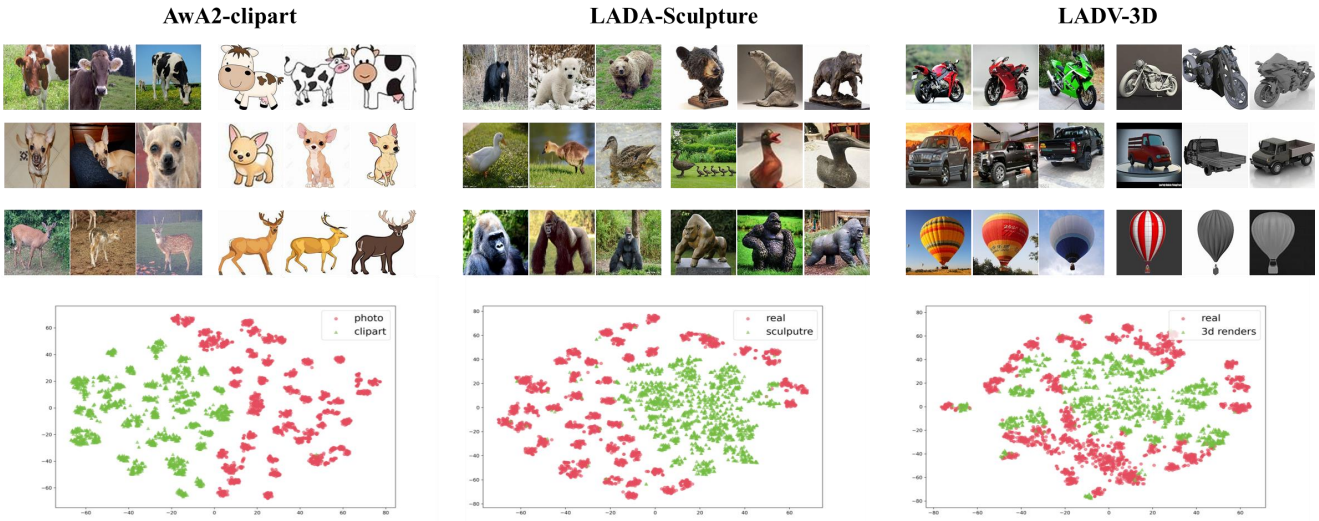


Figure 1. Some randomly selected samples and CLIP feature TSNE visualization of all samples on our proposed three benchmarks.

B. Empirical studies

This paper's key insights are derived from our empirical studies that the VLMs can interpret the visual domain shifts into language, in other words, visual domain shifts and descriptions of different domains are consistent in the VLM embedding space. The main paper lists the results between two unseen visual domains (*i.e*. "sketch" and "sculpture") and the training domain ("photo"). In this section, we list more results between unseen visual domains like "clipart", "3d model" and "painting" in Fig. 2. Similar to the main paper, the class-level descriptors generated by ConZIC [6] describe the visual difference. The style descriptions for the common photo domain (subtrahend) are generally implicit in the training caption corpus, thus, the class-level descriptors prominently consist of style descriptions about the unseen visual domains (minuend). To better visualize the global semantic direction, we aggregate all class-level descriptors into a final word cloud format, where prominent words represent the main visual direction between the two visual domains. For instance, semantical words "cartoon, view, character, draw, illustration" indicate the main style direction of the "clipart" domain. This demonstrates that the visual domain shifts can be approximated by some domain-related style descriptors, while these style descriptors can used to compute class-level textual domain differences.'

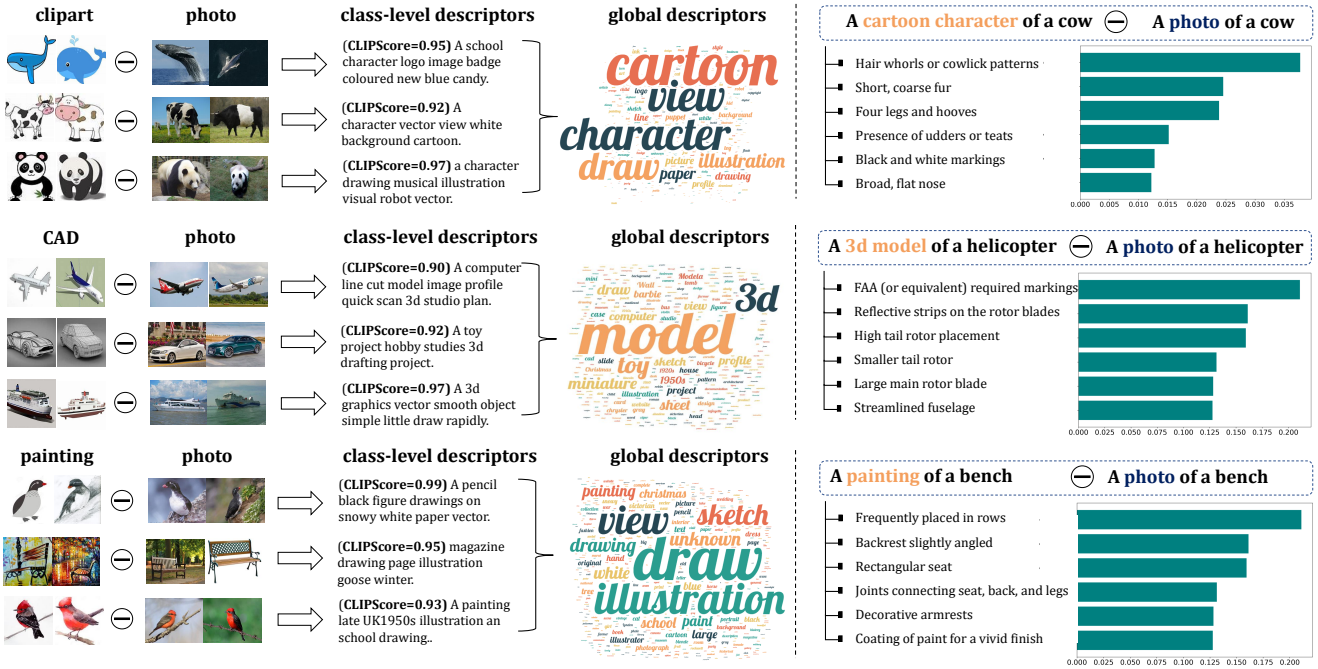
Then, we can utilize these textual domain differences to discover the domain-specific concepts. Concretely, we can concatenate these domain descriptors with specific classes in a prompt style, such as "A cartoon character of a cow", denotes as t_{tgt}. Similarly, we can construct a prompt for photo domains like "A photo of a cow", denoted as t_{src}. For each discriminative

visual concept c_i, we can compute the similarity s_i with this class-level textual domain shifts in the CLIP embedding space, as:

$$\Delta t = E_T(t_{tgt}) - E_T(t_{src}) \tag{1}$$

$$s_i = E_T(c_i) \cdot \Delta t, \tag{2}$$

where E_T means CLIP text encoder. Finally, we can get all concept activation $\{s_i\}_{i=1}^{M}$. We empirically find that those concepts with higher similarity are often domain-specific concepts that exhibit substantial variation between two domains. Therefore, $S = [s_1, ..., s_M]$ can viewed as a domain-specific concept activation a_{sp}. These observations provide the key insights of this paper.



(a) Interpret the visual domain shifts into language. (b) Discover the domain-specific concepts.

Figure 2. Empirical studies on more domain shifts.

C. More implementation details of DDO loss

In this section, we will introduce more implementation details about our proposed LanCE framework.

C.1. Generating domain descriptors

Based on the empirical observations in Sec. B, we aim to achieve generalization across a wide range of unseen visual domains by leveraging a large language model (LLM) to generate domain descriptors $\mathcal{P}$. A detailed list is shown in Fig. 3.

C.2. Textual domain shifts embeddings

After we get all domain descriptors $\mathcal{P} = \{p_i\}_{i=1}^{N_p}$, we can use these domain descriptors to compute textual domain shift embeddings based on Eq. (1). Specifically, we set the training domain as "photo" and compute the difference between two domain-related class prompts. Finally, we can get all class-level textual domain shifts $[\Delta t(p_i, y)]_{N_p \times N_y}$.

C.3. Details of DDO loss

DDO aims to encourage the orthogonality between all domain-specific concept activations $[\boldsymbol{a}_{sp}(p_i, y)]_{N_p \times N_y}$ and class prototype concept activations (*i.e.* linear weight W_F). The computation of $\boldsymbol{a}_{sp}(p_i, y)$ as:

$$\boldsymbol{a}_{sp}(p_i, y) = [\Delta t(p_i, y) \cdot E_T(c_1), ..., \Delta t(p_i, y) \cdot E_T(c_M)] \tag{3}$$

- ***Concept prompt:*** *What are useful visual features for identifying an {classname}. You can consider visual features from appearance, color, pattern, shape perspectives. please describe at least 50 visual features without class name in sentence list: [].*

- ***Domain Descriptor Prompt:*** *Generate 200 new domain description prompts modeled on "a photo of a {}"*

Answered Domain descriptors (processed): [painting, clipart, infographic, quickdraw, sketch, cartoon character, advertising posters, sculpture, watercolor, 3D model, blueprint, silhouette, vintage photo, digital art, pencil drawing, technical illustration, doodle, woodcut, lithograph, oil painting, gouache painting, embroidery, charcoal sketch, etching, tattoo design, comic strip, collage, pop art, manga style drawing, minimalist drawing, pixel art, sepia photo, retro style, low-poly model, fantasy art, surrealist painting, futuristic concept art, impressionist painting, hyperrealistic drawing, chalkboard sketch, photo collage, CGI render, stained glass, anime style, Renaissance painting, Victorian engraving, children's book illustration, pen and ink drawing, crayon drawing, blueprint diagram, steampunk version, street art mural, pastel drawing, scientific diagram, psychedelic art, editorial cartoon, pointillism painting, decorative ornament, folk art version, holographic image, VR model, bronze statue, ancient cave painting, sci-fi style, children coloring book, mixed media art, conceptual sketch, steampunk mechanical drawing, hand-painted mural, pastel artwork, illustrated diagram, vector graphic, medieval manuscript, baroque painting, Cubist painting, Art Deco version, Art Nouveau illustration, acrylic painting, neon sign, icon design, shadow silhouette, cut-paper art, tapestry, cross-stitch pattern, visual novel character, emoji, logo, banner, motion graphic, kinetic sculpture, vintage postcard, LED display, glass sculpture, sand art representation, flower arrangement, fabric pattern, Egyptian hieroglyph, 16-bit video game character, pottery design, metal engraving, origami model, cyberpunk illustration, graffiti stencil, stained glass panel, Rorschach inkblot, Gothic architecture detail, postage stamp, wireframe model, LEGO model, hologram, paper doll, bubble letter graffiti, cookie cutter shape, emoji sticker, flipbook animation, crystal carving, sand sculpture, totem pole, Moai statue, scientific model, photo negative, pop-up book, clay sculpture, fabric print, kinetic art piece, chalk pavement art, scrimshaw, augmented reality filter, laser-cut wood model, beadwork pattern, lenticular print, tarot card, astrological chart, glass mosaic, domino tile, rubber stamp, fashion illustration, tattoo flash, 2D animation cell, comic book panel, topographic map, ASCII art, street photography shot, stone carving, bookplate illustration, linocut, album cover, silhouette photo, flipbook, watercolor wash, 4-bit pixel icon, map illustration, animated GIF, 3D hologram, typography art, paper cutout, retrowave poster, constellation, steampunk icon, painted ceramic tile, abstract representation, shadow puppet, cave engraving, dot matrix print, Rube Goldberg machine, aerial view photo, album art design, topographic elevation map, needlepoint design, quilling art piece, badge design, marble statue, glass etching, logo badge, postage stamp illustration, embossed print, neon artwork, street poster, ancient rune, steampunk watch gear, environmental infographic, safety sign, blueprint schematic, wire sculpture, papercraft model, photorealistic painting, vintage label, linocut print, painting on driftwood, cave wall painting, tribal tattoo, doodle sticker, video game cover, emoji art, lava lamp pattern, comic character, floral arrangement shaped, retro poster, minimalist icon, classic movie poster, botanical illustration, cross-sectional diagram, video game avatar, medieval tapestry, carved pumpkin in the shape]

Figure 3. LLM prompts to generate visual concepts and domain descriptors and detailed generated domain descriptor list.

Model	FLOPs (in billions)	Parameters	Memory Usage (in MB)
baseline	13285.1	63224	918.0
LanCE	13287.7	63224	1038.1

Table 2. Comparison of computation complexity, including FLOPs, trainable parameters, and estimated memory usage.

Notably, the DDO loss is independent of specific input samples, as all $[\boldsymbol{a}_{sp}(p_i, y)]N_p \times N_y$ are processed by W_F collectively with each batch of image samples.

D. More ablation studies

Effect of CLIP backbone. Table. 3 and Table. 4 have shown the detailed results with CLIP ViT-B/32 and CLIP ViT-L/14, demonstrating that our proposed DDO regularizer can improve the OOD accuracy across different CLIP image backbones.

Effect of the numbers of domain descriptors. Fig. 4 provides the ablation studies on DomainNet. Similar to other results on other datasets, the OOD accuracy gradually improves with an increasing number of domain descriptors.

Effect of relevance of domain descriptors. Table. 5 has shown some relevant keywords about each benchmark. We remove domain descriptors containing these keywords and investigate the contribution of remaining domain-irrelevant domain descriptors to improving OOD accuracy. Fig. 6 shows the results on DomainNet.

Computation complexity. To evaluate the gained computation complexity brought by DDO loss, we list the comparison results of FLOPs, trainable parameters, and estimated memory usage. Results are shown in Table. 2, as we can see, the gained computation complexity is minor and almost negligible. It demonstrates that our proposed DDO loss is a plug-in loss that can be applied to many concept-based models without changing model architecture and increasing too much computation cost.

Model	Concept	Method	CUB-Painting ID	CUB-Painting OOD	AwA2-clipart ID	AwA2-clipart OOD	LADA-Sculpture ID	LADA-Sculpture OOD	LADV-3D ID	LADV-3D OOD
CLIP ViT-B/32										
CLIP ZS [2]	✗	✗	65.27	40.14	91.50	67.44	87.17	48.98	84.36	50.07
CLIP LP [2]			51.59	44.57	91.72	79.32	89.46	65.12	68.28	60.49
CLIP-CBM	human	baseline	61.05	35.08	92.38	62.14	94.52	48.89	87.85	53.25
		+DDO	**62.88**	**39.51**	**90.95**	**66.29**	**95.05**	**53.70**	**87.29**	**55.42**
PCBM† [5]	ConceptNet	baseline	58.31	38.83	93.41	**71.78**	95.31	55.64	90.11	56.87
		+DDO	**58.81**	**39.91**	**92.58**	69.11	**95.76**	**57.96**	**90.03**	**57.18**
LaBO [4]	LLM	baseline	67.57	35.35	93.90	62.60	96.70	77.63	99.44	56.00
		+DDO	**67.83**	**37.28**	**94.50**	**71.70**	**98.27**	**79.69**	**99.13**	**58.88**
CLIP ViT-L/14										
CLIP ZS [2]	✗	✗	62.21	52.77	95.70	90.26	91.26	82.05	71.82	66.29
CLIP LP [2]			82.00	61.40	97.11	86.75	96.81	74.40	93.68	63.81
CLIP-CBM	human	baseline	78.51	50.54	95.69	81.91	96.66	70.44	92.21	60.64
		+DDO	**78.70**	**55.53**	**95.71**	**83.72**	**96.77**	**75.76**	**92.59**	**63.51**
PCBM† [5]	ConceptNet	baseline	75.85	54.41	97.17	84.77	97.60	76.69	94.71	65.88
		+DDO	**76.48**	**57.50**	**97.19**	**86.58**	**97.64**	**79.74**	**94.82**	**68.33**
LaBO [4]	LLM	baseline	81.91	56.24	97.14	84.15	97.41	74.56	99.90	63.17
		+DDO	**82.34**	**59.60**	**97.26**	**87.66**	**98.12**	**80.00**	**99.93**	**68.01**

Table 3. Detailed accuracy performance comparison on single unseen domain benchmarks, including CUB-Painting, AwA2-clipart, LADA-Sculpture, and LADV-3D.

DomainNet.								
Model	Method	ID real	OOD clipart	OOD infograph	OOD painting	OOD quickdraw	OOD Sketch	OOD Avg
CLIP ViT-B/32								
LaBO [4]	baseline	86.11	60.60	35.00	54.64	8.40	49.7	41.67
	+DDO	**86.33**	**64.00**	**39.66**	**58.90**	**8.57**	**52.9**	**44.81**
CLIP ViT-L/14								
LaBO [4]	baseline	91.20	76.04	48.41	66.16	16.58	66.35	55.63
	+DDO	**91.29**	**77.37**	**53.00**	**68.91**	**17.27**	**69.04**	**56.20**

Table 4. Detailed accuracy performance comparison on multiple unseen domain benchmarks, *i.e*. DomainNet.

E. More qualitative results

Fig. 5 has shown more qualitative results about the top-5 visual concepts, ranked by the weights in W_F, demonstrating that our proposed DDO regularizer can reduce the correlation between domain-specific concepts and final predictions.

F. Human evaluation

To validate the efficacy of our proposed method, we conduct a human evaluation on the top 10 visual concepts that exhibit a high correlation with the final class, ranked by the weights W_F trained on DomainNet. The evaluation considers two key aspects: *Discriminability* and *Generalizability*. For each concept, we present several images from all visual domains and invite three human experts to assign a score ranging from 0 to 4. Specifically, for *Discriminability*, a score of 0 indicates the concept is unrelated to the corresponding category, while a score of 4 signifies the concept is a salient visual feature for

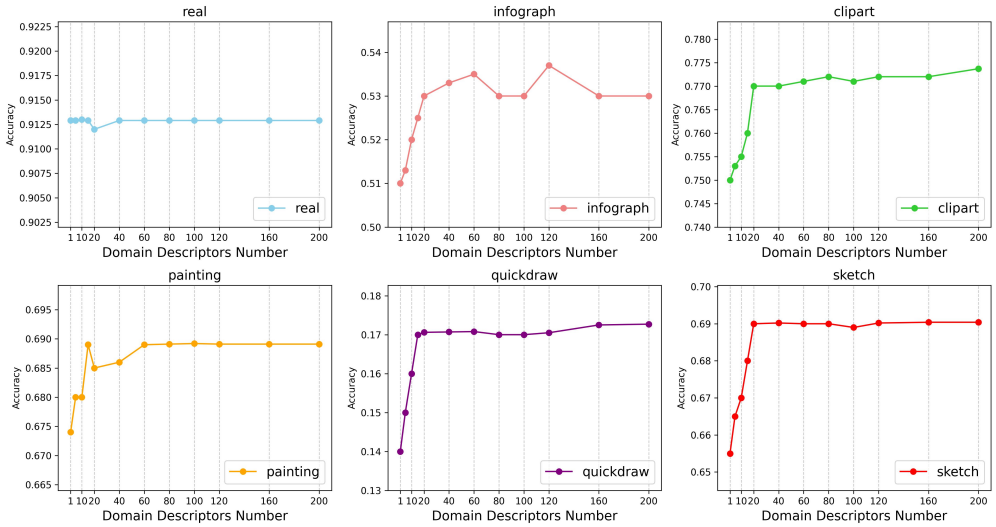


Figure 4. Ablation studies about the numbers of the domain descriptors on DomainNet.

Dataset	Relevant descriptors (keywords)
CUB-Painting	painting, sketch, watercolor, drawing, doodle, art
AwA2-clipart	clipart, cartoon, emoji, comic, anime, avatar, animated
LADA-Sculpture	sculpture, 3D, statue
LADV-3D	3D, CGI, VR, low-poly
DomainNet	painting, clipart, infographic, quickdraw, sketch, watercolor, cartoon, collage, art, drawing, sketch, illustration, doodle, poster emoji, comic, anime

Table 5. Relevant descriptors for each benchmark.

		real	clipart	painting	infograph	sketch	quickdraw
LanCE	baseline	91.20	76.04	66.16	48.41	66.35	16.58
	+DDO(IR)	91.20	76.60	67.80	50.00	67.74	16.30
	+DDO	**91.29**	**77.37**	**68.91**	**53.0**	**69.04**	**17.27**

Table 6. Ablation studies on DomainNet for the effect of relevance of the domain descriptors. +DDO(IR) only use the domain-irrlevant descriptors while +DDO use all domain descriptors.

the category. For *Generalizability*, a score of 0 indicates the concept exists only in a single domain, whereas a score of 4 represents a domain-invariant concept. The scores from the three annotators are averaged, and concepts with an average score greater than 2 are classified as either discriminative or domain-invariant concepts. For each concept, we generate a binary label based on these classifications. Finally, we analyze the percentage of discriminative and domain-invariant concepts to report the final results. To evaluate these two metrics, we select top-10 concepts for each class ranking by their weights in the

Model	Discriminability(%)	Generalizability(%)
baseline	75	64
LanCE	**79**	**82**

Table 7. Human evaluation about the percentage of distinguishing concepts and percentage of domain-invariant concepts.

final layer W_F, and ask annotators to judge whether each concept meets the demands above. The ratio of accurate concepts are shown in the Table. 7 where our proposed LanCE achieves better results than the baseline LaBO, demonstrating the effectiveness of languid-guided concept erasing design can significantly decrease the association between domain-specific concepts and the final output.

G. Limitations

Our method highly depends on pre-trained VLMs like CLIP and LLMs like GPT-3.5. However, these models are limited in application to some professional fields like medical treatments. We think further integration of an extra knowledge base and task-specific fine-tuning of these pre-trained models is a potential solution to solve these limitations. We hope this work can prompt the development of robust interpretable models.

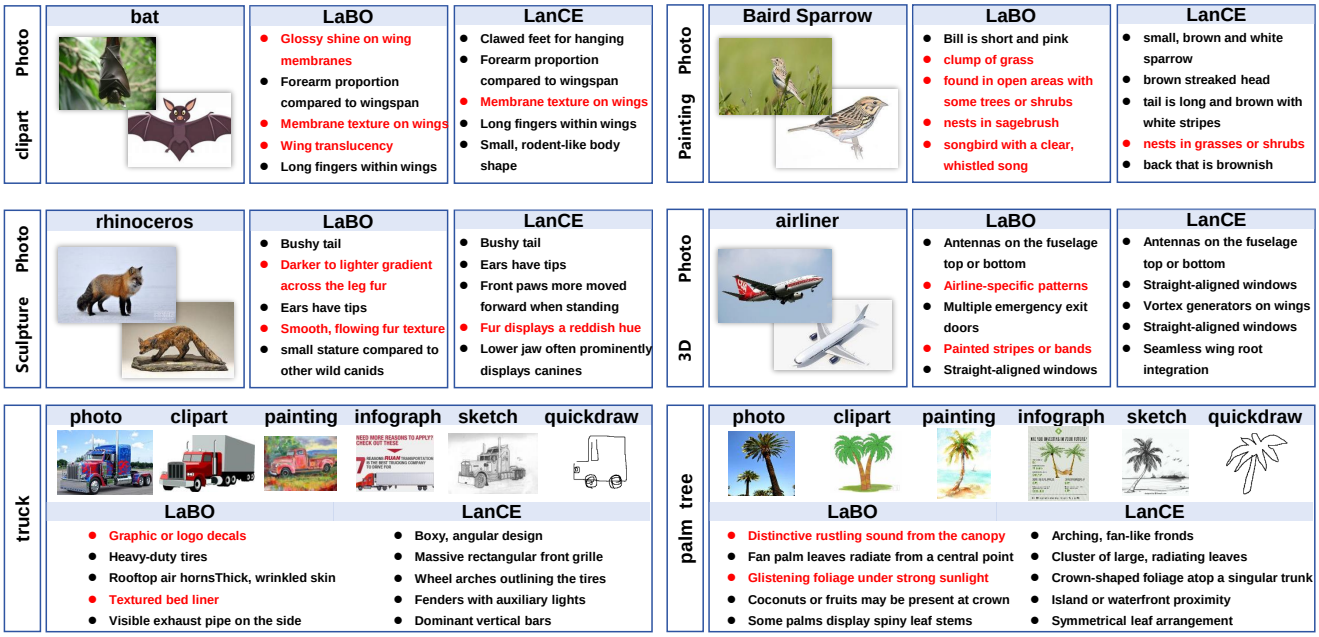


Figure 5. More qualitative results. top-5 concepts, ranking by their weights in the final linear layer WF . Baseline indicates the results of the original LaBo. Domain-specific concepts are hightlighted in red.

References

[1] Lisa Dunlap, Clara Mohri, Devin Guillory, Han Zhang, Trevor Darrell, Joseph E Gonzalez, Aditi Raghunathan, and Anna Rohrbach. Using language to extend to unseen domains. International Conference on Learning Representations (ICLR), 2023. 1

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[3] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. 1

[4] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023. 4

[5] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022. 4

[6] Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23465–23476, 2023. 1