The Neglected Tails in Vision-Language Models

Supplementary Material

Outline

This document supplements the main paper with comprehensive analyses and ablations. Below outlines the document.

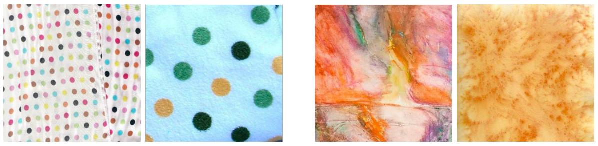
- **Section A.** We provide details of the nine benchmark datasets and the four ImageNet variants.
- **Section B.** We report our estimated concept frequency on the other eight benchmark datasets.
- **Section C.** We report REAL performance on head and tail classes across nine benchmark datasets.
- **Section D.** We attach all implementation details of REAL for reproducibility.
- **Section E.** We present further ablations of REAL-Linear to highlight the importance of synonym-based retrieval and cross-modal adaptation.
- **Section F.** We show that the performance gain of REAL can generalize across different architectures, pretraining datasets, and prompt templates.
- **Section G.** We show more failures of state-of-the-art multimodal systems (visual chatbots and text-to-image generative models) on diverse tailed concepts.
- **Section H.** We qualitatively show that REAL-Prompt can help generate images featuring rare concepts.

A. Dataset Details

Table 6 shows the details of the nine benchmarks, including the number of classes and the size of testset. These datasets are widely used in the research community of zero-shot recognition.

Table 6. **Details of thirteen benchmark datasets.**

Dataset	#Classes	#Testing data	Remark
Flowers [30]	102	2,463	flower classification
Cars [18]	196	8,041	car (brand and year) classification
Aircraft [26]	100	3,333	aircraft classification
Pets [33]	37	3,669	domestic pet classification
Food [5]	101	30,300	food classification
DTD [9]	47	1,692	texture classification
EuroSAT [13]	10	8,100	satellite imagery classification
CUB [43]	200	5,794	bird classification
ImageNet [10]	1,000	50,000	wordnet categories classification
ImageNet-V2 [17]	1,000	30,000	an ImageNet variant of temporal shift
ImageNet-A [15]	200	7,500	an ImageNet variant of adversarial samples
ImageNet-R [14]	200	30,000	an ImageNet variant of artistic renditions
ImageNet-S [46]	1,000	50,000	an ImageNet variant of sketches



(a) `dotted` vs. `polka dotted` (b) `smeared` vs. `stained`

Figure 6. **Classes in the DTD [9] dataset can be semantically ambiguous.** The texture class `dotted` is a super-set of another class `polka dotted`. For another case, people use the class name `smeared` and `stained` interchangeably.

B. Results of Concept Frequency Estimation

In Table 7, we plot the concept frequency calculated using our proposed method for the other eight benchmark datasets. Surprisingly, we find that all of them follow an imbalanced distribution (as measured in LAION). Moreover, we plot the per-class zero-shot accuracies grouped by concept frequency and confirm a strong correlation between concept frequency and zero-shot accuracy in the majority of the datasets except for the DTD dataset. For DTD, we find that certain classes can overlap with others. For example, the `dotted` and `polka dotted`, `smeared` and `stained` (see Figure 6). Such ambiguous labeling makes DTD an outlier for our frequency analysis.

C. Performance Breakdown of REAL

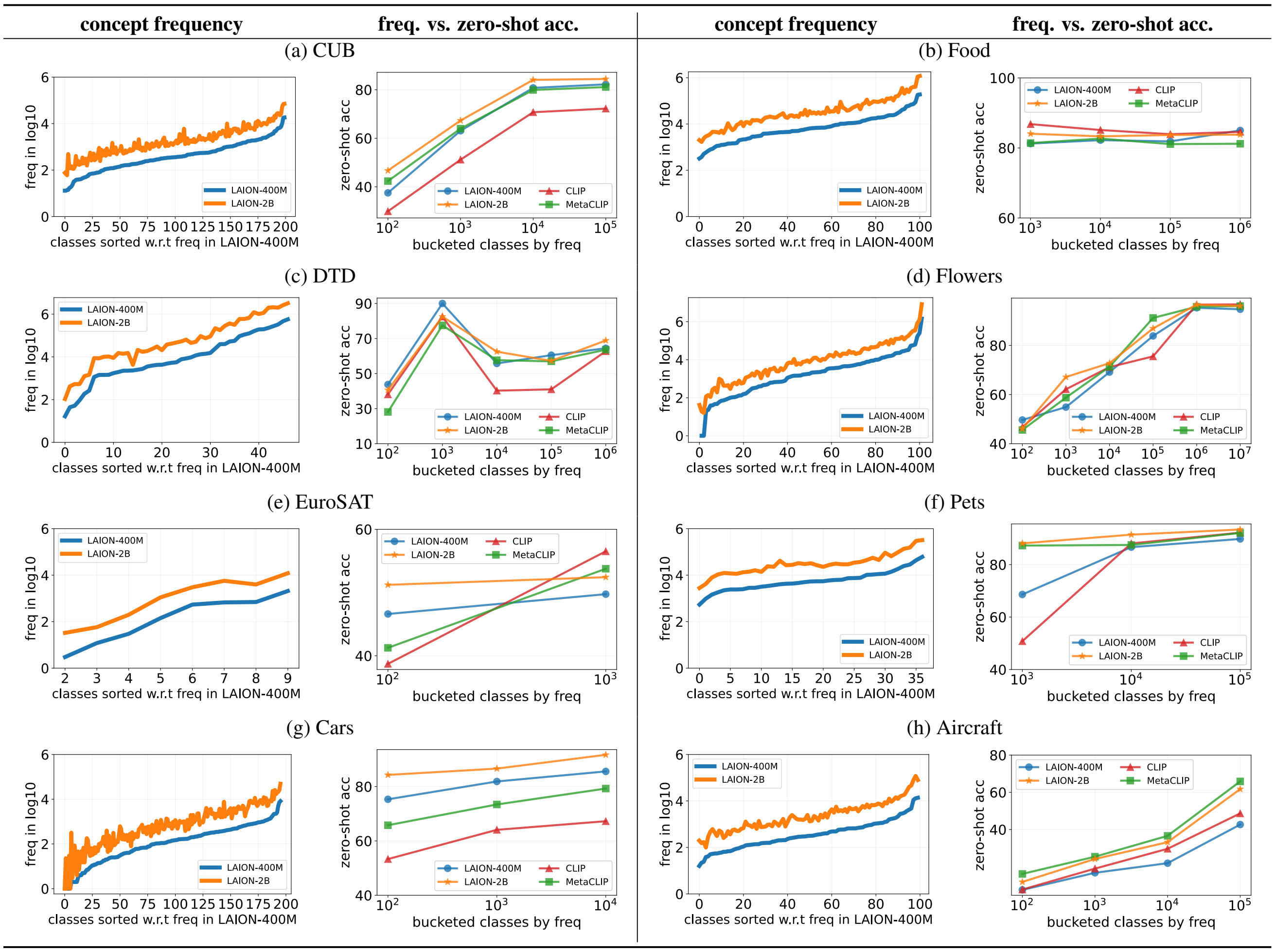
In Table 8, we show the improvement of REAL on the head and tail classes across nine benchmark datasets. We emphasize that REAL can significantly lift both head and tail accuracy on downstream tasks using the original pretraining data.

D. Further Details for REAL

Synonym filtering in REAL-Prompt. We use OpenCLIP's text encoder to filter out ChatGPT-generated synonyms that might be confused with other downstream concepts. Specifically, we retain only those synonyms that have the highest cosine similiarity scores with their original class names (not with another downstream concept). This filtering step is critical to REAL-Prompt's performance as shown in Table 15.

Linear probing in REAL-Linear. We follow previous work [22, 48] and adopt the same procedure and hyperparameters to learn a robust cross-modal classifier. Specifically, we initialize the weights of the cross-modal linear classifier using averaged text features constructed using the most frequent synonyms and OpenAI templates [35]. Next,

Table 7. **Vision-language models (VLMs) inherit long tails from their pretraining data.** We show that concepts from the other eight benchmark datasets all follow a long-tailed distribution in the pretraining datasets (e.g. LAION-400M [37], LAION-2B [38]). The strong correlation between concept frequency and accuracy prevalently exists among the datasets. For DTD, the trend deteriorates because of the ambiguous labeling of class names (see Figure 6).



(a) CUB

(b) Food

(c) DTD

(d) Flowers

(e) EuroSAT

(f) Pets

(g) Cars

(h) Aircraft

Table 8. **Breakdown improvements of REAL.** REAL-Prompt and REAL-Linear (500 retrieved examples per concept) can significantly improve upon the baseline performance using the OpenAI templates [35] on nine standard zero-shot recognition benchmark datasets. We define the tail as the 20% least frequent classes and the rest as the head for each dataset. REAL significantly lifts both head and tail accuracies on these datasets.

	Method	ImageNet		Flowers		Cars		Aircraft		CUB		Pets		Food		DTD		EuroSAT		Avg	
		Head	Tail	Head	Tail	Head	Tail	Head	Tail	Head	Tail	Head	Tail	Head	Tail	Head	Tail	Head	Tail	Head	Tail
LAION 400M	OpenAI templates	64.8	55.2	70.0	50.6	81.1	72.9	18.9	8.4	69.1	40.1	87.4	83.5	80.4	82.5	54.3	55.2	65.0	23.9	65.7	52.5
	REAL-Prompt	65.4	56.2	76.8	58.8	85.2	73.7	20.8	7.3	69.3	40.6	88.7	88.6	80.5	82.3	59.3	62.0	64.3	41.9	67.8	56.8
		+0.6	+1.0	+6.8	+8.2	+4.1	+0.8	+1.9	-1.1	+0.2	+0.5	+1.3	+5.1	+0.1	-0.2	+5.0	+6.8	-0.7	+18.0	+2.1	+4.3
	REAL-Linear	67.8	58.9	82.4	57.2	87.0	73.2	34.4	10.0	79.3	50.4	89.7	87.7	80.8	83.6	60.8	63.5	69.9	19.2	72.5	56.0
	(500)	+2.9	+3.7	+12.4	+6.6	+5.9	+0.3	+15.5	+1.6	+10.2	+10.3	+1.3	+4.2	+0.4	+1.1	+6.5	+8.3	+4.9	-4.7	+6.5	+3.5
LAION 2B	OpenAI templates	68.0	61.0	75.6	50.5	87.0	82.5	27.9	11.5	73.0	49.0	90.5	90.6	82.0	85.1	58.0	55.2	54.0	38.1	68.6	58.4
	REAL-Prompt	68.2	61.6	79.4	55.1	89.2	80.8	29.3	11.3	72.8	47.7	91.5	92.8	82.1	85.1	64.4	63.5	51.6	58.7	69.8	61.8
		+0.2	+0.6	+3.8	+4.6	+2.2	-1.7	+1.4	-0.2	-0.2	-1.3	+1.5	+2.2	+0.1	+0.0	+6.4	+8.3	-2.4	+20.6	+1.2	+3.4
	REAL-Linear	69.8	64.8	84.1	66.9	90.0	82.3	45.4	25.5	82.4	62.2	91.5	92.6	82.3	86.2	64.5	70.0	76.0	22.0	76.2	63.6
	(500)	+1.8	+3.8	+8.5	+16.4	+3.0	-0.2	+17.6	+14.0	+9.4	+13.2	+1.0	+2.0	+0.3	+1.1	+6.5	+14.8	+22.0	-16.1	+7.6	+5.2

we stick to the reported [22] learning rate of 1e-4 with a cosine annealing schedule, weight decay of 1e-2, batch size of 32, and training epochs of 10. Finally, we average the learned cross-modal classifier weights with the zero-shot classifier weights (as shown in Figure 4). We apply the same set of hyperparameters for all datasets and model architectures. We will release our code and retrieved data for reproducibility.

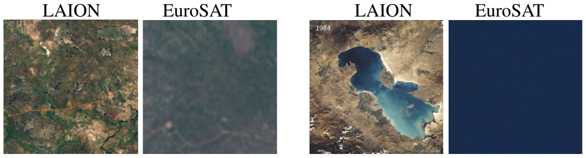
E. More Ablations of REAL-Linear

In this section, we show that synonym-based retrieval and cross-modal adaptation are crucial for the performance of REAL-Linear. We also explain the lower performance of REAL-Linear on Stanford Cars dataset. Lastly, we ablate the retrieval sizes v.s. zero-shot accuracies.

Synonyms help retrieve diverse data. It is crucial to retrieve images whose captions contain any of the concept synonyms instead of just the name predefined by the downstream task. Table 9 shows that using all synonyms can retrieve more diverse images for a performance boost of 4% from 64.2% to 68.2% when averaged across eight benchmark datasets, surpassing REACT Locked-Text's 65.5%. In addition, Table 10 shows that REAL-Linear outperforms another retrieval-augmented method Neural Priming [44], which does not consider concept synonyms for retrieval. For a fair comparison, we follow [44] to use 100 retrieved images per class because they do not release the models and hyperparameters.

Learning robust cross-modal classifiers. We show that cross-modal adaptation [22], which uses both text and image features to learn a linear classifier, is more robust against the distribution shifts between retrieved (pretraining) data and target domains. Concretely, Table 11 shows that performing naive linear probing using only retrieved images achieves lower accuracy by 6.4% averaged across all benchmark datasets, sometimes underperforms the zero-shot classifier [35] constructed using OpenAI prompt templates [35]. This shows that using both images and texts can effectively reduce overfitting to retrieved pretraining data.

Remarks on REAL-Linear's performance on the Stanford Cars dataset. Table 1 shows that the performance of our REAL-Linear on the Cars dataset [18] is 4% lower than that of REACT [24], despite that we relax the string matching criteria (by matching partial names) to retrieve more relevant images. We attribute the performance gap to the limited images retrieved from LAION-400M [37], owing to the fine-grained nature of the class names, e.g. "Audi S6 Sedan 2011". Supporting evidence is shown in Table 9, where using synonyms for retrieval increases the accuracy of Cars from 71.1% to 84.4%. This suggests future work on better retrieval methods for datasets with specific brand names.



(a) `Herbaceous Vegetation` (b) `Sea Lake`

Figure 7. **Large distribution shifts between LAION and EuroSAT.** We compare two randomly sampled satellite images from LAION and EuroSAT, for the class `Herbaceous Vegetation` and `Sea Lake`, respectively. Images from LAION present higher resolution and more distinct features while the EuroSAT images are blurry and lack informative features.

Remarks on retrieval sizes. Retrieving more pretraining examples generally helps REAL-Linear achieve higher accuracies for zero-shot recognition, as shown in Table 12. Yet, increasing the retrieval size from 100 to 500 per concept only improves accuracy by 0.9% (averaged over nine benchmarks). As such, we adopt 500 for our major experiments in this paper.

F. Generalization Performance of REAL

In this section, we show that REAL generalizes across model architectures, datasets, and prompt templates.

Generalizing across architectures and datasets. Table 13 shows that REAL-Linear consistently improves the zero-shot performance of OpenCLIP across different ViT architectures (B/32, B/16, and L/14) and LAION datasets (400M and 2B). Yet, both REAL-Linear and REACT [24] fail to improve on the EuroSAT dataset [13], presumably because satellite imagery is very rare in LAION (e.g., we can retrieve at most one image for `Annual Crop` and `Herbaceous Vegetation`). In addition, the few retrieved satellite images in LAION are drastically different from EuroSAT testset images due to sensor shifts, as shown in Figure 7.

Generalizing across prompt templates. Table 14 shows that REAL-Prompt is effective regardless of the prompt templates (OpenAI [35], DCLIP [28], and CuPL [34]).

G. More Failures of Multimodal Systems

In Figure 8 and 9, we show more failure cases of state-of-the-art multimodal systems on tailed concepts identified by our frequency estimation method. These tailed concepts are randomly sampled from nine benchmark datasets and span across a variety of domains, including birds, flowers, fungi, snakes, frogs, fish, household items, and more. We qualitatively test the visual recognition abilities of two most popular visual chatbots: GPT-4V [50] (trained on proprietary data) and LLaVA1.5 [23] (trained on open-source data using a frozen CLIP image encoder). We also test the image generation abilities of two most popular generative models:

Table 9. **Using concept synonyms helps retrieve more diverse pretraining images.** Retrieving images whose captions contain any of the concept synonyms (instead of just the name predefined by the downstream task) can improve the performance of REAL-Linear (using 500 images per class). We attach the performance of REACT Locked-Text (using 10K images per class) for reference.

Method	Images per class	ImageNet	Flowers	Cars	Aircraft	Pets	Food	DTD	EuroSAT	**Avg**
REACT Locked-Text	10K	65.7	73.1	**88.5**	24.5	89.2	81.8	49.8	51.1	65.5
REAL-Linear (without synonyms)	500	64.6	72.6	71.1	27.9	88.6	81.9	56.3	50.8	64.2
REAL-Linear (with synonyms)	500	**65.9**	**78.8**	84.4	**29.6**	**89.5**	**81.4**	**61.5**	**51.5**	**67.8**

Table 10. **REAL-Linear outperforms Neural Priming**. We compare REAL-Linear with Neural Priming [44] using the ViT-B/16 model pre-trained on LAION-2B [38]. REAL-Linear consistently outperforms Neural Priming on all their reported benchmarks, presumably because Neural Priming does not consider synonyms for retrieval.

Method	Images per class	ImageNet	Cars	Flowers	Aircraft	Food	Pets	**Avg**
Neural Priming	100	70.8	89.3	79.8	33.0	86.7	91.9	75.3
REAL-Linear	100	**71.9**	**90.3**	**81.9**	**38.7**	**86.7**	**92.2**	**77.0**

DALL-E 3 [41] (trained on proprietary data) and Stable Diffusion XL [2] (trained on open-source data using a frozen CLIP text encoder). We observe that these systems fail to recognize or generate more than half of the tailed concepts we sampled. In particular, LLaVA1.5 and Stable Diffusion XL fail on all these tailed concepts, suggesting a large performance gap between proprietary and open-source multimodal systems.

H. REAL-Prompt for Generative Models

Figure 10 and 11 contains qualitative results of REAL-Prompt on state-of-the-art text-to-image generative models including DALL-E 3 [41] and Stable Diffusion XL [2]. This shows that using most frequent synonyms can help generate correct images for tailed concepts. We also note that our method is more effective on the more capable generative model DALL-E 3, presumably because it is trained with more data than open-source Stable Diffusion XL. This suggests opportunities for future work to improve open-source VLMs on image synthesis for rare concepts.

Table 11. **Cross-modal adaptation improves the robustness of REAL-Linear.** We highlight that using both images and texts during training can help address the distribution shifts between pretraining data and target domains. Concretely, we adopt cross-modal WiSE-FT [22, 48], which first learns a cross-modal linear classifier using both retrieved image features and text features constructed using the most frequent concept synonyms and OpenAI templates [35]. This cross-modal classifier is then ensembled with a zero-shot classifier whose weights are text features of the most frequent synonyms averaged across OpenAI prompt templates. We show that this cross-modal strategy is much more robust against vanilla image-only linear probing that uses only retrieved image features, which overfits to retrieved data and sometimes underperforms the zero-shot classifier.

Method	ImageNet	Flowers	Cars	Aircraft	CUB	Pets	Food	DTD	EuroSAT	**Avg**
OpenAI templates [35]	62.9	68.0	79.2	16.7	63.8	86.7	80.9	54.5	51.5	62.7
REAL-Linear (image-only)	62.1$^{-0.8}$	78.0$^{+10.0}$	77.5$^{-1.7}$	33.1$^{+16.4}$	73.1$^{+9.3}$	86.1$^{-0.6}$	79.5$^{-1.4}$	53.8$^{-0.7}$	15.6$^{-35.9}$	62.4$^{-0.3}$
REAL-Linear (cross-modal)	65.9$^{+3.0}$	78.8$^{+10.8}$	84.1$^{+7.9}$	29.6$^{+12.9}$	74.0$^{+10.2}$	89.5$^{+2.8}$	81.4$^{+0.5}$	61.5$^{+6.2}$	51.5$^{+0.0}$	67.8$^{+6.1}$

Table 12. **Zero-shot accuracy vs. retrieval size.** We conducted an ablation study the impact of retrieval size for REAL-Linear, and for comparison, we included results using OpenAI templates and our REAL-Prompt. Notably, even with a smaller retrieval size of 100 images per concept, we achieve strong performance (only 1% lower on avg.), though our best results come with a retrieval size of 500 images per concept.

Number of shots	ImageNet	Flowers	Cars	Aircraft	CUB	Pets	Food	DTD	EuroSAT	**Avg**
OpenAI templates [35]	62.9	68.0	79.2	16.7	63.8	86.7	80.9	54.5	51.5	62.7
REAL-Prompt	63.6	76.6	82.7	18.0	64.0	88.8	81.0	59.9	57.5	65.8$^{+3.1}$
REAL-Linear (100)	65.3	77.8	84.0	25.1	72.4	89.3	81.0	60.4	53.3	67.6$^{+4.9}$
REAL-Linear (500)	65.9	78.8	84.4	29.6	74.0	89.5	81.4	61.5	51.5	68.5$^{+5.8}$

Table 13. **REAL-Linear generalizes across different pretraining datasets and architectures.** REAL-Linear consistently achieves performance gains with three OpenCLIP architectures (ViT B/32, B/16, and L/14) and pretraining datasets (LAION 400M and 2B). For reference, we attach the performance REACT reported on these benchmarks. Notably, our REAL-Linear (500 examples per class) even outperforms REACT Gated-Image (10K examples per class) when both use a larger visual encoder ViT-L/14. We highlight the **best accuracy** in bold and underline the $\underline{\text{second best}}$ numbers for ImageNet.

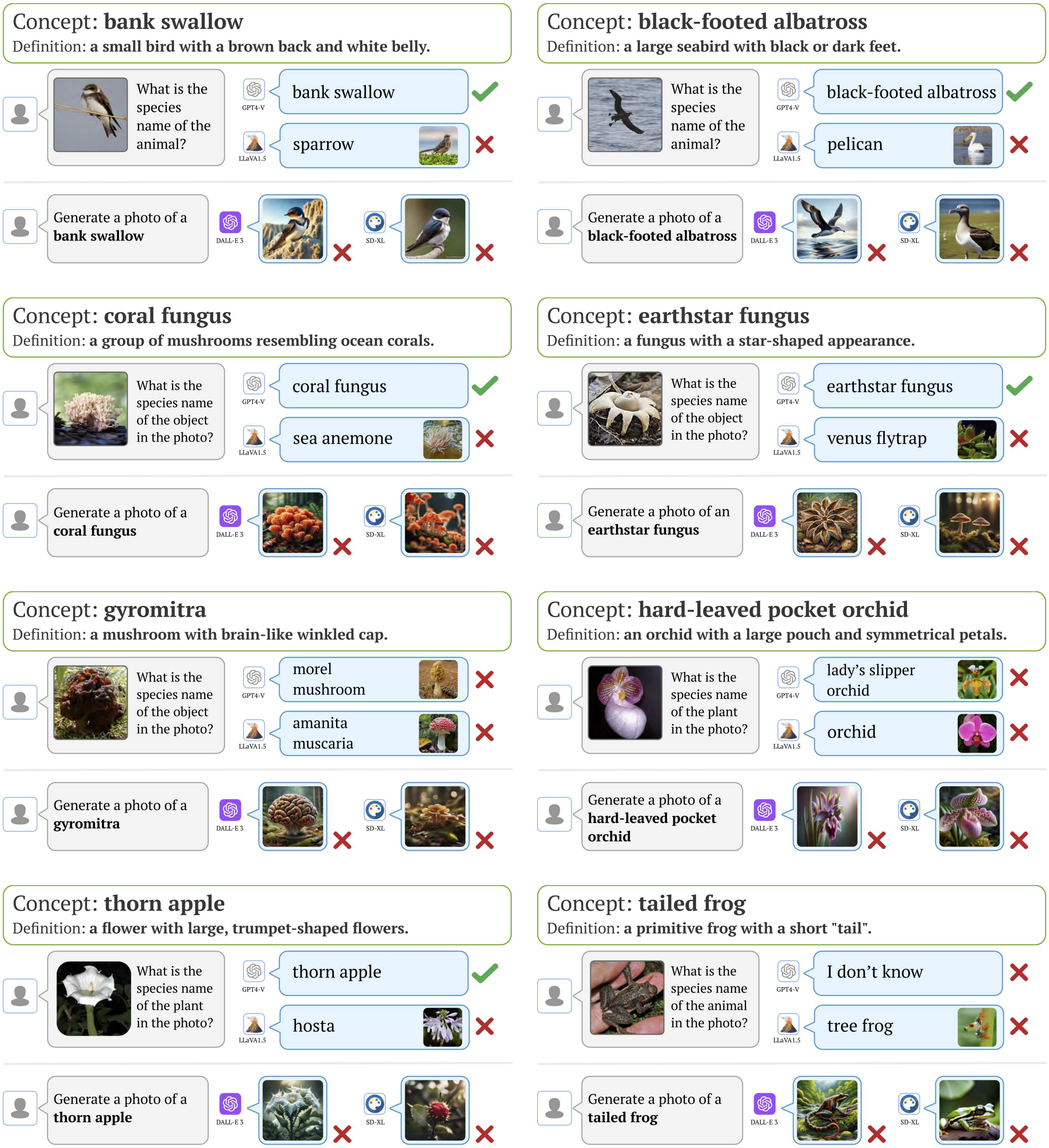
	Arch	Method	ImageNet	Flowers	Cars	Aircraft	CUB	Pets	Food	DTD	EuroSAT	Avg
LAION 400M	ViT B/32	OpenAI templates [35]	62.9	68.0	79.2	16.7	63.8	86.7	80.9	56.1	51.5	62.6
		REACT Locked-Text (10K)	$\underline{65.7}$	73.1	88.5	24.5	–	89.2	81.8	49.8	51.1	–
		REACT Gated-Image (10K)	64.2	72.3	88.1	24.8	–	89.5	83.0	51.4	45.4	–
		REAL-Linear (500)	**65.9**	78.8	84.1	29.6	74.0	89.5	81.4	61.5	51.5	68.5
	ViT B/16	OpenAI templates [35]	67.0	69.2	83.6	17.7	67.2	89.3	86.2	51.3	50.3	64.6
		REACT Locked-Text (10K)	69.9	–	–	–	–	–	–	–	–	–
		REACT Gated-Image (10K)	**70.5**	–	–	–	–	–	–	–	–	–
		REAL-Linear (500)	$\underline{69.6}$	80.6	86.5	31.5	79.1	91.3	86.4	61.4	51.9	71.0
	ViT L/14	OpenAI templates [35]	$\underline{72.7}$	75.4	89.5	24.9	76.4	91.8	90.0	60.2	62.3	71.5
		REAL-Linear (500)	**74.4**	85.4	91.0	40.2	84.8	93.4	90.3	66.5	59.8	76.2
LAION 2B	ViT B/32	OpenAI templates [35]	66.6	71.8	86.0	24.5	68.5	90.6	82.7	56.1	48.0	66.1
		REACT Locked-Text (10K)	67.5	–	–	–	–	–	–	–	–	–
		REACT Gated-Image (10K)	**69.6**	–	–	–	–	–	–	–	–	–
		REAL-Linear (500)	$\underline{68.8}$	80.6	88.4	41.3	78.5	91.7	83.1	65.6	51.9	72.2
	ViT B/16	OpenAI templates [35]	$\underline{70.2}$	71.4	88.2	26.9	72.7	90.5	86.5	56.3	53.4	68.5
		REAL-Linear (500)	**72.4**	83.4	90.3	45.6	83.6	92.2	87.1	66.0	46.9	74.2
	ViT L/14	OpenAI templates [35]	75.3	75.2	91.9	36.6	78.5	93.2	91.0	62.8	64.6	74.3
		REACT Gated-Image (10K)	$\underline{76.5}$	–	–	–	–	–	–	–	–	–
		REAL-Linear (500)	**76.9**	86.5	92.6	55.3	87.5	94.7	91.2	69.4	57.9	79.1

Table 14. **REAL-Prompt generalizes across prompt templates.** We show that REAL-Prompt (using the most frequent synonyms) can improve upon both OpenAI prompt templates [35] and LLM-enriched templates such as DCLIP [28] and CuPL [34].

	Arch	Method	ImageNet	Flowers	Cars	Aircraft	CUB	Pets	Food	DTD
LAION-400M	ViT-B/32	OpenAI templates [35]	62.9	68.0	79.2	16.7	63.8	86.7	80.9	54.5
		+ REAL-Prompt	63.6	76.6	82.7	18.0	64.0	88.8	81.0	59.9
		DCLIP [28]	62.1	–	–	–	64.5	84.6	80.1	51.4
		+ REAL-Prompt	62.9	–	–	–	64.7	88.1	80.0	55.5
		CuPL [34]	63.7	65.8	80.0	17.8	–	87.4	79.5	59.1
		+ REAL-Prompt	64.2	72.3	81.7	18.3	–	88.0	79.5	59.3
	ViT-B/16	OpenAI templates [35]	67.0	69.2	83.6	17.7	67.2	89.3	86.2	51.0
		+ REAL-Prompt	67.6	77.1	84.4	19.5	67.3	91.0	86.3	58.1
		DCLIP [28]	65.8	–	–	–	68.6	86.2	85.2	51.1
		+ REAL-Prompt	66.2	–	–	–	68.6	89.8	85.2	57.1
		CuPL [34]	67.8	67.9	83.4	18.6	–	89.7	85.2	57.9
		+ REAL-Prompt	68.1	73.1	84.0	18.8	–	90.5	85.2	59.8
	ViT-L/14	OpenAI templates [35]	72.7	75.4	89.5	24.9	76.4	91.8	90.0	60.2
		+ REAL-Prompt	72.9	82.9	89.9	26.0	76.4	93.3	90.2	63.6
		DCLIP [28]	71.8	–	–	–	77.2	89.2	89.3	57.7
		+ REAL-Prompt	72.3	–	–	–	77.3	92.1	89.4	60.5
		CuPL [34]	73.3	76.9	89.3	27.5	–	92.4	89.4	65.4
		+ REAL-Prompt	73.7	82.4	89.6	28.2	–	92.8	89.4	65.7
LAION-2B	ViT-B/32	OpenAI templates [35]	66.6	71.8	86.0	24.5	68.5	91.8	82.7	57.4
		+ REAL-Prompt	66.9	76.2	87.5	25.6	68.2	91.8	82.7	64.2
		DCLIP [28]	65.7	–	–	–	68.5	90.5	81.2	53.2
		+ REAL-Prompt	66.0	–	–	–	68.2	90.6	81.2	57.7
		CuPL [34]	67.0	69.5	86.5	26.5	–	91.0	81.6	62.7
		+ REAL-Prompt	67.3	74.1	87.6	27.4	–	91.1	81.6	63.8
	ViT-B/16	OpenAI templates [35]	70.2	71.4	88.2	26.9	72.7	91.6	86.5	57.9
		+ REAL-Prompt	70.3	78.6	88.7	28.7	72.6	91.7	86.6	64.8
		DCLIP [28]	69.5	–	–	–	73.6	91.6	86.0	58.1
		+ REAL-Prompt	69.7	–	–	–	73.5	91.7	86.0	62.7
		CuPL [34]	70.6	70.6	88.6	29.6	–	91.1	86.2	63.8
		+ REAL-Prompt	70.8	76.6	89.4	30.0	–	91.1	86.2	64.9
	ViT-L/14	OpenAI templates [35]	75.3	75.2	91.9	36.6	78.5	94.1	91.0	64.1
		+ REAL-Prompt	75.4	83.4	92.1	37.6	78.5	94.2	91.0	67.8
		DCLIP [28]	74.5	–	–	–	78.3	93.2	90.8	63.1
		+ REAL-Prompt	74.9	–	–	–	78.2	93.2	90.8	64.4
		CuPL [34]	75.7	75.4	92.6	41.2	–	94.3	90.6	68.7
		+REAL-Prompt	75.9	82.0	92.1	41.4	–	94.2	90.6	68.8

Table 15. **The importance of synonym-filtering for REAL-Prompt.** After obtaining synonyms from ChatGPT, we use OpenCLIP's text encoder to filter the synonyms that might be confused with other downstream concepts. We show that this filtering step is critical for REAL-Prompt's performance.

	ImageNet	Flowers	Cars	Aircraft	CUB	Pets	Food	DTD
REAL-Prompt w/o Synonym Filtering	50.5	45.0	59.3	9.5	55.6	39.9	63.5	10.9
REAL-Prompt w/ Synonym Filtering	63.6	76.6	82.7	18.0	64.0	88.8	81.0	59.9



Concept: **bank swallow**
Definition: **a small bird with a brown back and white belly.**
What is the species name of the animal?
bank swallow ✓
sparrow ✗
Generate a photo of a **bank swallow** ✗ ✗

Concept: **black-footed albatross**
Definition: **a large seabird with black or dark feet.**
What is the species name of the animal?
black-footed albatross ✓
pelican ✗
Generate a photo of a **black-footed albatross** ✗ ✗

Concept: **coral fungus**
Definition: **a group of mushrooms resembling ocean corals.**
What is the species name of the object in the photo?
coral fungus ✓
sea anemone ✗
Generate a photo of a **coral fungus** ✗ ✗

Concept: **earthstar fungus**
Definition: **a fungus with a star-shaped appearance.**
What is the species name of the object in the photo?
earthstar fungus ✓
venus flytrap ✗
Generate a photo of an **earthstar fungus** ✗ ✗

Concept: **gyromitra**
Definition: **a mushroom with brain-like winkled cap.**
What is the species name of the object in the photo?
morel mushroom ✗
amanita muscaria ✗
Generate a photo of a **gyromitra** ✗ ✗

Concept: **hard-leaved pocket orchid**
Definition: **an orchid with a large pouch and symmetrical petals.**
What is the species name of the plant in the photo?
lady's slipper orchid ✗
orchid ✗
Generate a photo of a **hard-leaved pocket orchid** ✗ ✗

Concept: **thorn apple**
Definition: **a flower with large, trumpet-shaped flowers.**
What is the species name of the plant in the photo?
thorn apple ✓
hosta ✗
Generate a photo of a **thorn apple** ✗ ✗

Concept: **tailed frog**
Definition: **a primitive frog with a short "tail".**
What is the species name of the animal in the photo?
I don't know ✗
tree frog ✗
Generate a photo of a **tailed frog** ✗ ✗

Figure 8. **State-of-the-art multimodal systems fail to recognize or generate tailed concepts (part 1).** We show more failure cases of popular multimodal systems (GPT-4V [50], LLaVA1.5 [23], DALL-E 3 [41], and Stable Diffusion XL [2]) on tailed concepts sampled from standard benchmark datasets such as ImageNet [10], Flowers [30], Aircraft [26], and etc. For GPT-4V and LLaVA1.5, we include example images of incorrectly predicted classes to show that visual chatbots often misclassify rare concepts as some similar-looking yet more common concepts. We include a definition for each tailed concept to show that DALL-E 3 and Stable Diffusion (SD-XL) can fail to capture the correct colors, shapes, and other characteristics of these concepts.

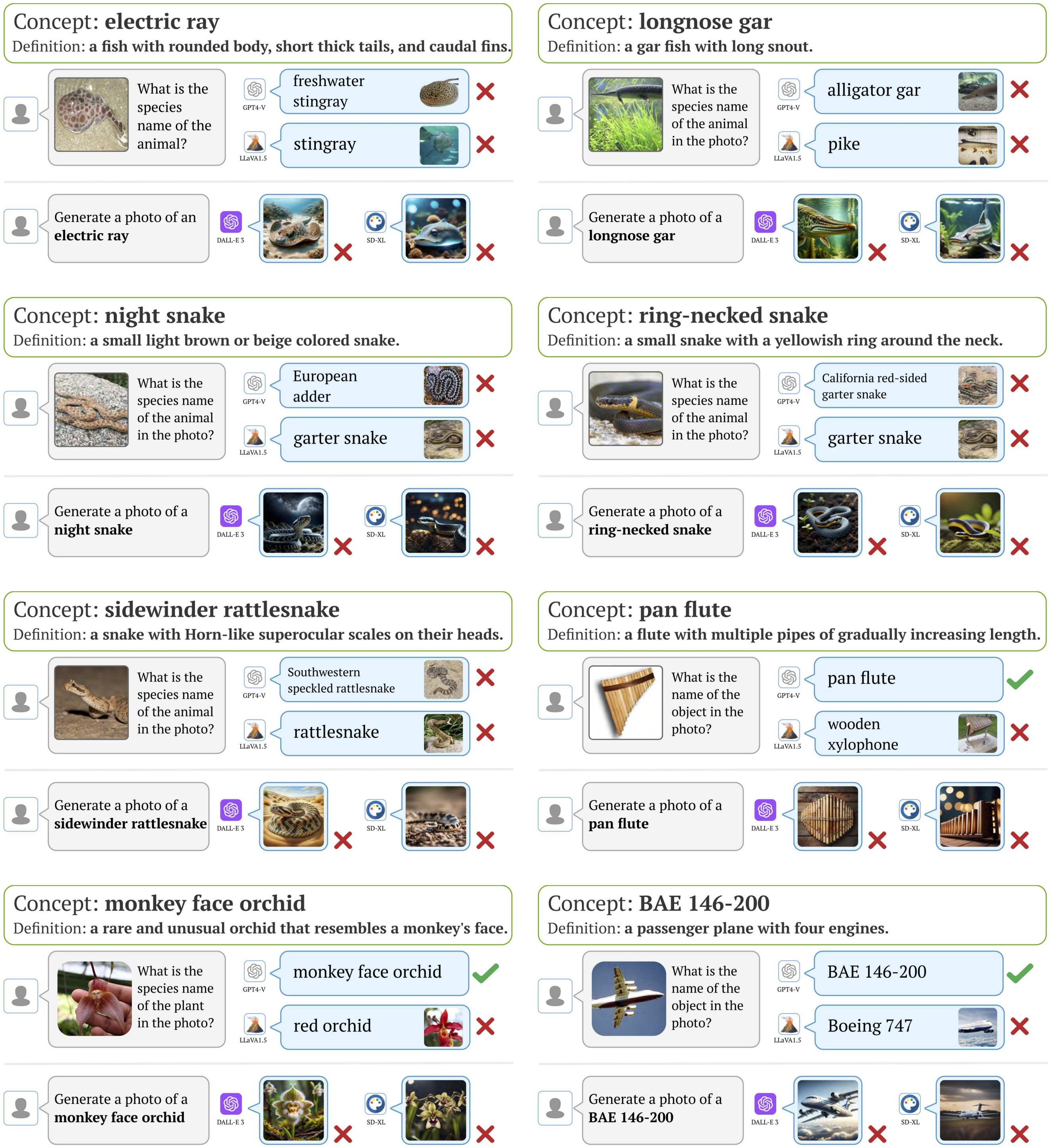
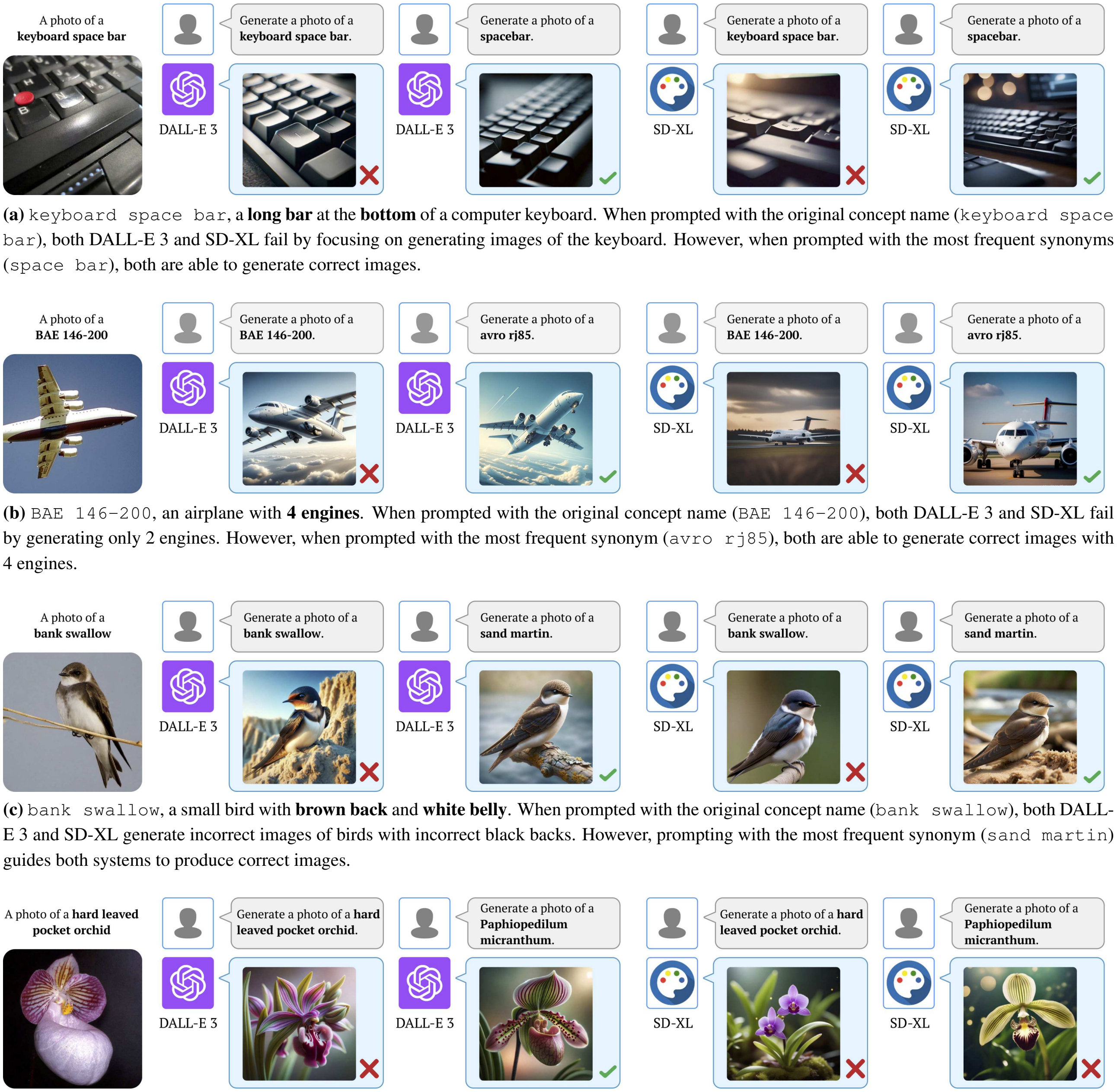

Concept: electric ray
Definition: a fish with rounded body, short thick tails, and caudal fins.
What is the species name of the animal?
GPT4-V
freshwater stingray ✗
LLaVA1.5
stingray ✗
Generate a photo of an electric ray
DALL-E 3 ✗
SD-XL ✗
Concept: longnose gar
Definition: a gar fish with long snout.
What is the species name of the animal in the photo?
GPT4-V
alligator gar ✗
LLaVA1.5
pike ✗
Generate a photo of a longnose gar
DALL-E 3 ✗
SD-XL ✗
Concept: night snake
Definition: a small light brown or beige colored snake.
What is the species name of the animal in the photo?
GPT4-V
European adder ✗
LLaVA1.5
garter snake ✗
Generate a photo of a night snake
DALL-E 3 ✗
SD-XL ✗
Concept: ring-necked snake
Definition: a small snake with a yellowish ring around the neck.
What is the species name of the animal in the photo?
GPT4-V
California red-sided garter snake ✗
LLaVA1.5
garter snake ✗
Generate a photo of a ring-necked snake
DALL-E 3 ✗
SD-XL ✗
Concept: sidewinder rattlesnake
Definition: a snake with Horn-like superocular scales on their heads.
What is the species name of the animal in the photo?
GPT4-V
Southwestern speckled rattlesnake ✗
LLaVA1.5
rattlesnake ✗
Generate a photo of a sidewinder rattlesnake
DALL-E 3 ✗
SD-XL ✗
Concept: pan flute
Definition: a flute with multiple pipes of gradually increasing length.
What is the name of the object in the photo?
GPT4-V
pan flute ✓
LLaVA1.5
wooden xylophone ✗
Generate a photo of a pan flute
DALL-E 3 ✗
SD-XL ✗
Concept: monkey face orchid
Definition: a rare and unusual orchid that resembles a monkey's face.
What is the species name of the plant in the photo?
GPT4-V
monkey face orchid ✓
LLaVA1.5
red orchid ✗
Generate a photo of a monkey face orchid
DALL-E 3 ✗
SD-XL ✗
Concept: BAE 146-200
Definition: a passenger plane with four engines.
What is the name of the object in the photo?
GPT4-V
BAE 146-200 ✓
LLaVA1.5
Boeing 747 ✗
Generate a photo of a BAE 146-200
DALL-E 3 ✗
SD-XL ✗

Figure 9. **State-of-the-art multimodal systems fail to recognize or generate tailed concepts (part 2).** We show more failure cases of popular multimodal systems (GPT-4V [50], LLaVA1.5 [23], DALL-E 3 [41], and Stable Diffusion XL [2]) on tailed concepts sampled from standard benchmark datasets such as ImageNet [10], Flowers [30], Aircraft [26], and etc. For GPT-4V and LLaVA1.5, we include example images of incorrectly predicted classes to show that visual chatbots often misclassify rare concepts as some similar-looking yet more common concepts. We include a definition for each tailed concept to show that DALL-E 3 and Stable Diffusion (SD-XL) can fail to capture the correct colors, shapes, and other characteristics of these concepts.


A photo of a keyboard space bar
Generate a photo of a keyboard space bar.
DALL-E 3
Generate a photo of a spacebar.
DALL-E 3
Generate a photo of a keyboard space bar.
SD-XL
Generate a photo of a spacebar.
SD-XL
A photo of a BAE 146-200
Generate a photo of a BAE 146-200.
DALL-E 3
Generate a photo of a avro rj85.
DALL-E 3
Generate a photo of a BAE 146-200.
SD-XL
Generate a photo of a avro rj85.
SD-XL
A photo of a bank swallow
Generate a photo of a bank swallow.
DALL-E 3
Generate a photo of a sand martin.
DALL-E 3
Generate a photo of a bank swallow.
SD-XL
Generate a photo of a sand martin.
SD-XL
A photo of a hard leaved pocket orchid
Generate a photo of a hard leaved pocket orchid.
DALL-E 3
Generate a photo of a Paphiopedilum micranthum.
DALL-E 3
Generate a photo of a hard leaved pocket orchid.
SD-XL
Generate a photo of a Paphiopedilum micranthum.
SD-XL

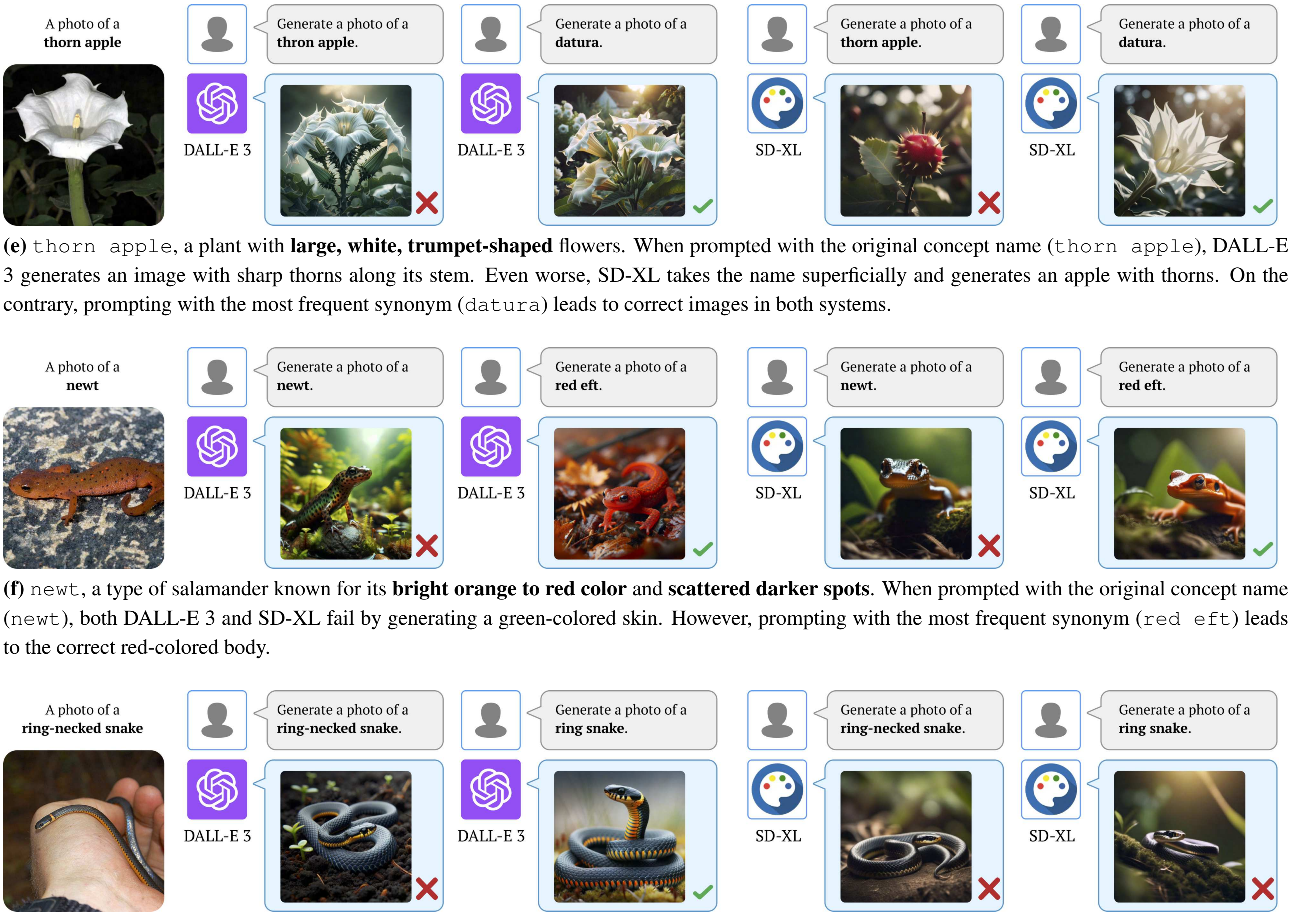
(a) `keyboard space bar`, a **long bar** at the **bottom** of a computer keyboard. When prompted with the original concept name (`keyboard space bar`), both DALL-E 3 and SD-XL fail by focusing on generating images of the keyboard. However, when prompted with the most frequent synonyms (`space bar`), both are able to generate correct images.

(b) `BAE 146-200`, an airplane with **4 engines**. When prompted with the original concept name (`BAE 146-200`), both DALL-E 3 and SD-XL fail by generating only 2 engines. However, when prompted with the most frequent synonym (`avro rj85`), both are able to generate correct images with 4 engines.

(c) `bank swallow`, a small bird with **brown back** and **white belly**. When prompted with the original concept name (`bank swallow`), both DALL-E 3 and SD-XL generate incorrect images of birds with incorrect black backs. However, prompting with the most frequent synonym (`sand martin`) guides both systems to produce correct images.

(d) `hard leaved pocket orchid`, a type of orchid with a **distinctive pouch** and **symmetrical large petals**. When prompted with the original concept name (`hard leaved pocket orchid`), both DALL-E 3 and SD-XL generate incorrect images (note the missing pocket and shape of the petals). However, when prompted with the most frequent synonym (`Paphiopedilum micranthum`), DALL-E 3 produces the correct image. In contrast, SD-XL is able to recover the shape of petals but still misses the pocket.

Figure 10. **Prompting with the most frequent synonym can help DALL-E 3 and Stable Diffusion generate correct images (part 1).** We show more examples when DALL-E 3 [41] and Stable Diffusion XL (SD-XL) [2] initially fail to generate correct images for tailed concepts when prompted with their original concept names in standard classification benchmark datasets. We sample diverse tail concepts covering a variety of domains including household items, birds, flowers, insects, reptiles, and etc. We show that REAL-Prompt (prompting with the most frequent synonyms) often helps DALL-E 3 and Stable Diffusion produce correct images. We notice that for the `hard leaved pocket orchid` and `ring-necked snake`, the generated images of SD-XL improve but are still inaccurate. This suggests future work to improve open-source generative models on rare concepts.


A photo of a thorn apple
Generate a photo of a thron apple.
DALL-E 3
Generate a photo of a datura.
DALL-E 3
Generate a photo of a thorn apple.
SD-XL
Generate a photo of a datura.
SD-XL
A photo of a newt
Generate a photo of a newt.
DALL-E 3
Generate a photo of a red eft.
DALL-E 3
Generate a photo of a newt.
SD-XL
Generate a photo of a red eft.
SD-XL
A photo of a ring-necked snake
Generate a photo of a ring-necked snake.
DALL-E 3
Generate a photo of a ring snake.
DALL-E 3
Generate a photo of a ring-necked snake.
SD-XL
Generate a photo of a ring snake.
SD-XL

(e) `thorn apple`, a plant with **large, white, trumpet-shaped** flowers. When prompted with the original concept name (`thorn apple`), DALL-E 3 generates an image with sharp thorns along its stem. Even worse, SD-XL takes the name superficially and generates an apple with thorns. On the contrary, prompting with the most frequent synonym (`datura`) leads to correct images in both systems.

(f) `newt`, a type of salamander known for its **bright orange to red color** and **scattered darker spots**. When prompted with the original concept name (`newt`), both DALL-E 3 and SD-XL fail by generating a green-colored skin. However, prompting with the most frequent synonym (`red eft`) leads to the correct red-colored body.

(g) `ring-necked snake`, a small snake with a **yellowish ring around the neck**. When prompted with the original concept name (`ring-necked snake`), both DALL-E 3 and SD-XL fail by missing the yellow ring around the snake's neck. However, prompting with the most frequent synonym (`ring snake`) helps DALL-E 3 recover the ring. Meanwhile, SD-XL still fails to capture the ring which is likely due to insufficient relevant images in its pretraining data.

Figure 11. **Prompting with the most frequent synonym can help DALL-E 3 and Stable Diffusion (SD-XL) generate correct images (part 2).** We show more examples when DALL-E 3 [41] and Stable Diffusion XL (SD-XL) [2] initially fail to generate correct images for tailed concepts when prompted with their original concept names in standard classification benchmark datasets. We sample diverse tail concepts covering a variety of domains including household items, birds, flowers, insects, reptiles, etc. We show that REAL-Prompt (prompting with the most frequent synonyms) often helps DALL-E 3 and Stable Diffusion produce correct images. We notice that for the `hard-leaved pocket orchid` and `ring-necked snake`, the generative images of SD-XL improve but are still inaccurate. This suggests future work to improve open-source generative models on rare concepts.