A. Search procedure

In this work, we use combinatorial testing (CT) [38, 4] for searching systematic errors in the operational design domain. CT selects the set of subgroups to be tested a priori without taking the classifier's loss on subgroups into account when selecting the next subgroup. This has the advantage of providing certain coverage guarantees with respect to the ODD, e.g., pairwise testing ($n_C = 2$) will test every combination of two attributes at least once. Moreover, CT also allows evaluating multiple loss functions concurrently. However, a potential disadvantage is that CT does not search explicitly for subgroups of maximal loss and might thus be less efficient compared to targeted search procedures such as evolutionary algorithms (EAs). We compare CT with an EA in this section. The EA uses a population size of 25, tournament selection with tournament size 3, mutation probability 0.6, crossover probability 0.3, and resampling probability 0.1. In a mutation, a single semantic dimension is reset to a random value from the possible values of the respective dimension. In a crossover, values from the two genotypes are selected randomly with equal probability.

Figure 9 compares CT with this EA on the Vehicle Experiment from Section 5. Here, EA maximizes the risk for the "pickup" target class. We observe that EA outperforms CT for $n_C \in \{3, 4, 5\}$ on the risk for the pickup target class, but with a relatively small difference. However, EA performs subpar for other target classes, which indicates that every target class would require a separate EA run per target class. In contrast, a single run of CT performs reasonable well on all target classes. In summary, we use CT in this paper because it allows assessing the risk of multiple target classes concurrently within a single run.

B. ImageNet Experiments

B.1. Experimental Setting: Vehicle Experiment

We evaluate the following models with weights for image classification on ImageNet1k from torchvision [36]: VGG16 [47], ResNet50 [21], ConvNeXt-B [32], ViT-B/16 [12], and ViT-L/32 [12]. We focus on a subset of classes belonging to the vehicle subcategory, more specifically on misclassifying samples of the class "minivan" $\tilde{\mathbf{y}} = \mathbf{y}_{minivan}$ into other classes that have a distance of 2 in the WordNet [14] hierarchy:

- amphibian, amphibious vehicle (id: 408)
- fire engine, fire truck (id: 555)
- garbage truck, dustcart (id: 569)
- go-kart (id: 573)
- golfcart, golf cart (id: 575)
- moving van (id: 675)
- pickup, pickup truck (id: 717)
- police van, police wagon, paddy wagon, patrol wagon, wagon, black Maria (id: 734)
- snowplow, snowplough (id: 803)
- tow truck, tow car, wrecker (id: 864)
- trailer truck, tractor trailer, trucking rig, rig, articulated lorry, semi (id: 867)

We exclude classes with a WordNet distance of 1 since their visual appearance might be very similar to a "minivan" and our focus is not on fine-grained misclassifications.

We focus on an operational design domain $\mathbf{Z}$ with five semantic dimensions with the following values:

- *viewpoint*: center, side, front, rear
- object *size*: "", small, large, huge
- object *color*: "", black, white, gray, red, green, blue, yellow, orange, purple, magenta, cyan, brown
- *weather*: "", rainy, snowy, lightning, foggy, sunny
- *background*: background, forest, desert, lake, mountain, beach, city, river, house, tree, field, lawn, garden, street, people

The first of the possible values corresponds to a neutral choice, by which a specific dimension is not controlled. We observed that this can be preferable if a dimension is not relevant and leaving it empty simplifies the prompt for the text-to-image model. We use the prompt template

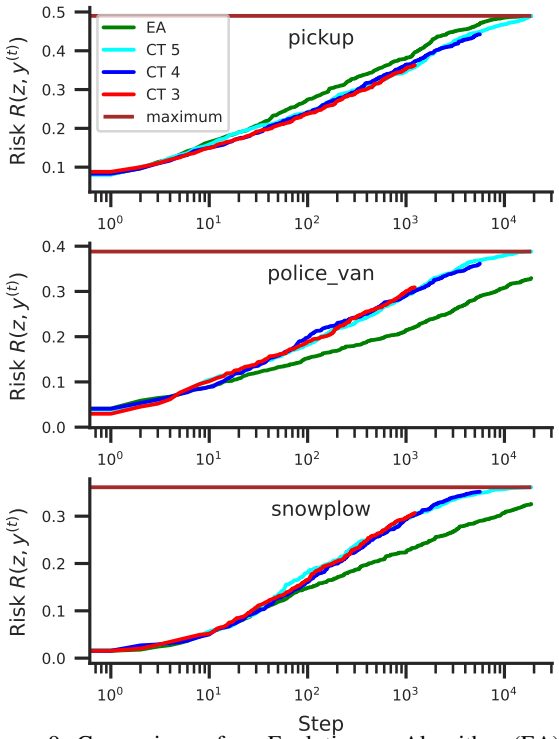

Figure 9. Comparison of an Evolutionary Algorithm (EA) and Combinatorial n_C-wise Testing (CT n_C). Step refers to the number of subgroups explored, lines show highest risk of any subgroup tested thus far, averaged over 100 repetitions. "Maximum" refers to the highest risk of any subgroup in the operational domain.

T_p ="{viewpoint} view of {size} {color} (minivan:1.5) in front of {weather} {background}". We use combinatorial testing with $n_C = 3$, exploring $|\mathbf{Z}_C| = 1.230$ out of $|\mathbf{Z}| = 4 * 4 * 13 * 6 * 15 = 18.720$ subgroups, and generate $n_S = 16$ image samples per subgroup using Stable Diffusion v1.5. We employ allpairspy [1] for combinatorial testing. See Table 1 for detailed results.

B.2. Experimental Setting: Person Experiment

We evaluate the following models with weights for image classification on ImageNet21k from timm [52]: MLP-Mixer-B/16 and MLP-Mixer-L/16 [49] . We focus on misclassifying samples of the class "homo" $\tilde{\mathbf{y}} = \mathbf{y}_{homo}$ (id: 3574) into the class "ape" (id: 3569). We skip logits corresponding to all other classes (some of which might be larger than the ones for homo and ape) and thus analyze effectively a hypothetical binary classifier derived from the pretrained 21k-class models without any finetuning.

We focus on an operational design domain $\mathbf{Z}$ with five semantic dimensions with the following values:

- *age*: "", young, old
- *gender*: "", female, male
- geographic *region*: "", european, american, hispanic, russian, arab, chinese, indian, african, australian
- *hairtype*: "", curly, short, long, blond, black, red, brown, gray
- *background*: background, forest, desert, lake, mountain, beach, city, river, house, tree, field, lawn, garden, street, people

The first of the possible values corresponds to a neutral choice, by which a specific dimension is not controlled. We observed that this can be preferable if a dimension is not relevant and leaving it empty simplifies the prompt for the text-to-image model. We use the prompt template T_p ="A {age} {gender} {region} (person:1.5) with {hairtype} hairs in front of {background}". We use combinatorial testing with $n_C = 3$, exploring $|\mathbf{Z}_C| = 1.371$ out of $|\mathbf{Z}| = 3*3*10*9*15 = 12.150$ subgroups, and generate $n_S = 16$ image samples per subgroup using Stable Diffusion v1.5. We employ allpairspy [1] for combinatorial testing. See Table 2 for detailed results.

C. Samples Zero-Shot Benchmark

We illustrate samples obtained for different hyperparameter settings that were quantitatively evaluated as part of the zero-shot systematic error benchmark (see Section 4.2). Figure 10 illustrates samples for different versions of Stable Diffusion [43]. Figure 11 illustrates samples for different number of steps n_t of the DPMSolver++ [33, 34]. Figure 12 illustrates samples for different prompt class weights w_c in the prompt template T_p ="An image of a *color type* (car:w_c) with a *background* background.".

D. Samples ImageNet Experiments

We illustrate 30 samples of source class "minivan" misclassified as "snowplow" (Figure 13), "pickup" (Figure 14), and "police van" (Figure 15). Moreover, we illustrate 30 samples of source class "person" misclassified as "ape" (Figure 16).

(Target) class	viewpoint	size	color	weather	background	$R(\mathbf{z}, \mathbf{y}^{(t)})$
ConvNeXt-B						
minivan	front	-	-	sunny	people	0.436
amphibian	center	small	brown	foggy	river	0.066
moving_van	front	huge	blue	rainy	garden	0.093
pickup	front	-	-	sunny	people	0.270
police_van	front	-	black	rainy	street	0.140
snowplow	front	huge	purple	snowy	field	0.129
ViT-L/32						
minivan	front	-	-	sunny	people	0.332
amphibian	side	huge	black	-	river	0.096
moving_van	rear	large	yellow	foggy	field	0.216
pickup	front	-	-	sunny	people	0.328
police_van	front	huge	yellow	lightning	street	0.177
snowplow	rear	small	orange	snowy	forest	0.285
ViT-B/16						
minivan	front	-	black	rainy	people	0.283
amphibian	center	small	red	foggy	river	0.124
moving_van	rear	large	yellow	foggy	garden	0.202
pickup	front	-	red	lightning	people	0.288
police_van	front	-	black	rainy	people	0.151
snowplow	rear	small	orange	snowy	forest	0.275
ResNet50						
minivan	front	-	-	sunny	people	0.597
amphibian	center	small	brown	foggy	river	0.065
moving_van	center	small	yellow	snowy	street	0.095
pickup	front	-	-	sunny	people	0.250
police_van	front	large	green	lightning	house	0.323
snowplow	center	large	black	snowy	street	0.187
VGG16						
minivan	rear	small	yellow	rainy	city	0.583
amphibian	center	-	orange	foggy	beach	0.118
moving_van	front	huge	yellow	sunny	house	0.195
pickup	front	-	-	sunny	people	0.344
police_van	rear	small	yellow	rainy	city	0.293
snowplow	rear	small	orange	snowy	forest	0.317
Averaged over models						
minivan	front	-	-	sunny	people	0.408
amphibian	center	small	brown	foggy	river	0.066
moving_van	front	huge	blue	rainy	garden	0.093
pickup	front	-	-	sunny	people	0.270
police_van	front	-	black	rainy	street	0.140
snowplow	front	huge	purple	snowy	field	0.129

Table 1. Detailed results for the "Vehicle Experiment" discussed in Section 5. We summarize systematic errors for source class $\tilde{\mathbf{y}}$="minivan" (higher $R(\mathbf{z})$ corresponding to stronger error) and systematic misclassifications into $\mathbf{y}^{(t)} \in$ {"amphibian", "moving_van", "pickup", "police_van", "snowplow"} (higher $R(\mathbf{z}, \mathbf{y}^{(t)})$ corresponding to stronger misclassifications). For each of the 5 studied models as well as averaged over all models, we show the subgroup corresponding to the strongest systematic error/misclassification and the corresponding risk R. The three highlighted lines correspond to the subgroups shown in Figure 1. Overall, identified subgroups differ considerably across models.

(Target) class	age	gender	region	hairtype	background	$R(\mathbf{z}, \mathbf{y}^{(t)})$
			Mixer-B/16			
ape	old	male	african	long	background	0.44462
ape	old	female	hispanic	red	tree	0.37845
ape	old	male	african	black	mountain	0.35407
ape	old	male	african	red	background	0.32113
ape	old	male	african	curly	garden	0.31325
ape	old	male	african	-	people	0.29942
ape	old	female	european	curly	tree	0.29639
ape	old	-	african	-	city	0.29433
ape	old	female	-	gray	people	0.29029
ape	old	male	african	curly	people	0.27311
ape	young	-	european	short	desert	0.00031
ape	young	male	hispanic	brown	desert	0.00031
ape	young	female	european	curly	desert	0.00028
ape	young	female	-	curly	desert	0.00027
ape	young	-	hispanic	blond	desert	0.00027
			Mixer-L/16			
ape	young	female	arab	brown	field	0.29910
ape	old	female	arab	gray	tree	0.27421
ape	young	female	indian	long	tree	0.20752
ape	old	female	arab	blond	house	0.18673
ape	young	female	arab	brown	tree	0.17659
ape	young	female	arab	brown	lawn	0.17200
ape	young	female	arab	gray	house	0.16944
ape	young	female	arab	short	lawn	0.16491
ape	-	-	australian	brown	people	0.15520
ape	young	female	arab	short	tree	0.15072
ape	-	female	hispanic	curly	street	0.00046
ape	-	-	hispanic	blond	street	0.00036
ape	young	-	hispanic	gray	city	0.00034
ape	young	male	-	curly	street	0.00034
ape	young	-	hispanic	black	street	0.00027

Table 2. Detailed results for the "Person Experiment" discussed in Section 5. We summarize systematic misclassifications into $\mathbf{y}^{(t)}$ ="ape" (higher $R(\mathbf{z}, \mathbf{y}^{(t)})$ corresponding to stronger misclassifications). For both studied models, we show the 10 subgroups corresponding to the top-ranked systematic misclassifications and the corresponding risk R as well as 5 subgroups where $R \approx 0$. We note that the two models have distinctive but different patterns in their top-ranked subgroups: An MLP-Mixer-B/16 [49] has several subgroups with high risk for "old male african" persons. An MLP-Mixer-L/16 [49] has several subgroups with high risk for "young female arab" persons. Moreover, the MLP-Mixer-L/16 has generally lower risk $R(\mathbf{z}, \mathbf{y}^{(t)})$) among the top-ranked subgroups.

Figure 10. Samples for different versions of Stable Diffusion (SD) [43]. We observe that SD v1-5 results in samples with good attribute binding while for SD 2-base and SD 2-1-base, object colour leaks into the background. Moreover, objects sometimes exhibit only partially the specified colour, while larger parts are dyed in other colours such as white (specifically for vans) for SD 2-base and SD 2-1-base. SD v1-5 does not exhibit this issue. This explains the better performance of PROMPTATTACK with SD v1-5 in Section 4.2. The 8 samples from left to right were generated for the prompts:
"an image of a green van (car:1.0) with a mountain background."
"an image of a blue van (car:1.0) with a desert background."
"an image of a blue cabriolet (car:1.0) with a desert background."
"an image of a green sedan (car:1.0) with a beach background."
"an image of a black van (car:1.0) with a mountain background."
"an image of a black van (car:1.0) with a beach background."
"an image of a black cabriolet (car:1.0) with a desert background."
"an image of a red SUV (car:1.0) with a forest background."

Figure 11. Samples for different number of steps n_t of the DPMSolver++ [33, 34]. As expected, more steps correspond to more realistic samples. However, even with $n_t = 5$ steps, PROMPTATTACK is able to reliably identify systematic errors (see Section 4.2). The 8 samples from left to right were generated for the prompts:
"an image of a green van (car:1.0) with a mountain background."
"an image of a blue van (car:1.0) with a desert background."
"an image of a blue cabriolet (car:1.0) with a desert background."
"an image of a green sedan (car:1.0) with a beach background."
"an image of a black van (car:1.0) with a mountain background."
"an image of a black van (car:1.0) with a beach background."
"an image of a black cabriolet (car:1.0) with a desert background."
"an image of a red SUV (car:1.0) with a forest background."

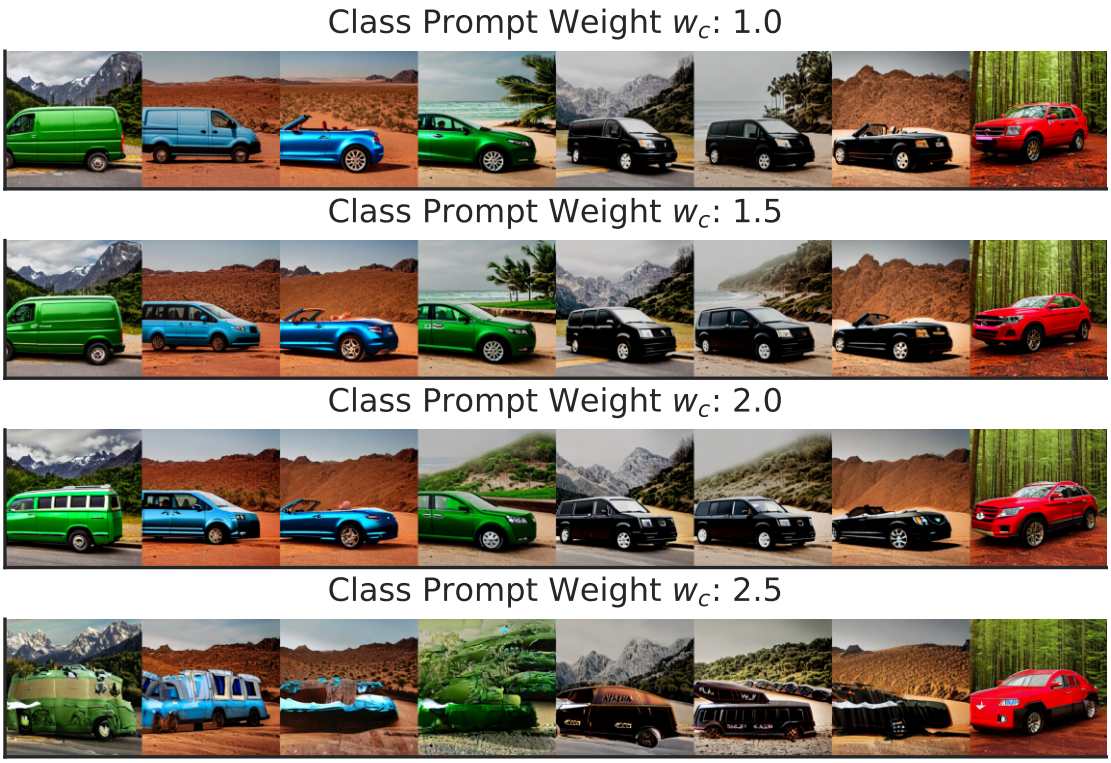

Figure 12. Samples for different prompt class weight w_c for the prompt template T_p ="An image of a *color type* (car:w_c) with a *background* background.". The improved performance of PROMPTATTACK for $w_c = 1.5$ and $w_c = 1.5$ compared to $w_c = 1.0$ is difficult to attribute to apparent visual properties of the samples. However, for $w_c = 2.5$, visual quality of samples strongly deteriorates, explaining the worse performance of PROMPTATTACK for this choice. The 8 samples from left to right were generated for the prompts:
"an image of a green van (car:w_c) with a mountain background."
"an image of a blue van (car:w_c) with a desert background."
"an image of a blue cabriolet (car:w_c) with a desert background."
"an image of a green sedan (car:w_c) with a beach background."
"an image of a black van (car:w_c) with a mountain background."
"an image of a black van (car:w_c) with a beach background."
"an image of a black cabriolet (car:w_c) with a desert background."
"an image of a red SUV (car:w_c) with a forest background."

Figure 13. 30 samples from prompt “rear view of small orange (minivan:1.5) in front of snowy forest.” that are misclassified as snowplows by a VGG16. Please note that actual viewpoints are a mix or “side” and “rear” views, and not purely “rear” views.

Figure 14. 30 samples from prompt “front view of (minivan:1.5) in front of sunny people.” that are misclassified as pickups by a ViT-L/32 [12]. Please note that often, there are no “people” in the background, indicating a shortcoming in the text-to-image model.

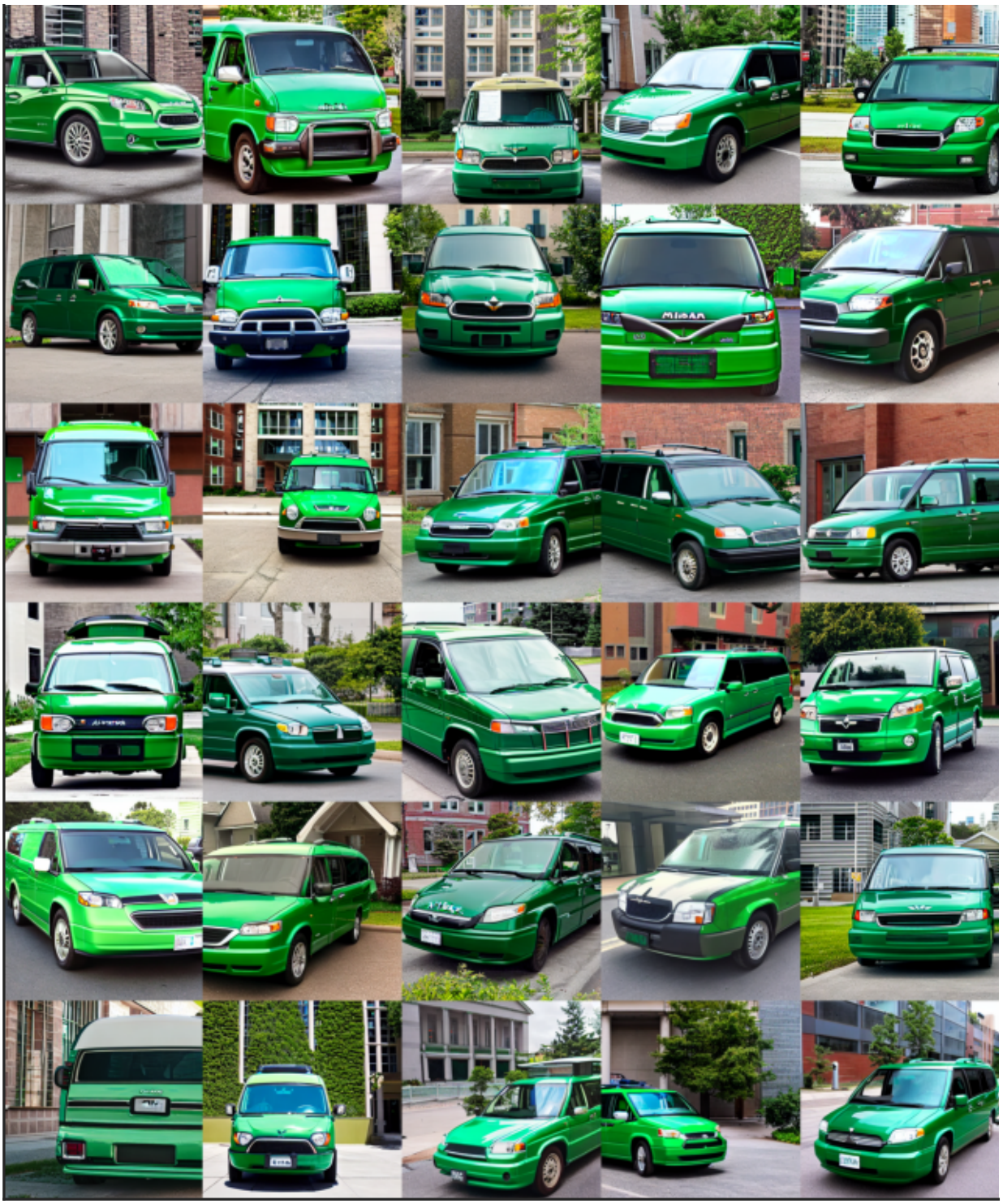

Figure 15. 30 samples from prompt “front view of large green (minivan:1.5) in front of lightning house.” that are misclassified as police-vans by a ResNet50. Please note that “lightning” is typically interpreted as a well illuminated scene and not as an actually lightning.

Figure 16. 30 samples from prompt “A old male african (person:1.5) with long hairs in front of background” that get a higher score for ape than for homo by a MLP-Mixer-B/16 [49] trained on ImageNet21k. We note that the samples from the text-to-image model are relatively similar and not fully representative of “old male african persons with long hairs”; this systematic error thus presumably correspond to a narrower subgroup than specified by above prompt.