A. Appendix

A.1. Assets

We provide a list of the assets used in our work (datasets, code, and models) in Table 2 (links) and Table 3 (licences).

A.2. Datasets

To design transformation queries from ImageNet classes, we have grouped classes into clusters by semantic similarity, upon manual inspection of the WordNet hierarchy of classes. The resulting clusters are shown in Table 8. This process resulted in 273 classes gathered in 47 clusters. We have not included all ImageNet classes because (i) we wanted to reduce the large number of dog breed classes, and (ii) a lot of classes were "standalone classes" with no natural target for transformation among the other classes. The clusters are then grouped into 13 bigger "groups" that are used solely for visualization in Figure 6 of the main paper.

A.3. Evaluation metrics

LPIPS. As recommanded by the authors of [58], we use the AlexNet [37] backbone to compute LPIPS distance, when we use it as an evaluation metric. To avoid using the same metric for optimization, we compute the LPIPS in the perceptual regularization term $\mathcal{L}_{perc}$, see Eq. (3) of the main paper, using the VGG16 network [37].

The LPIPS distance is computed at an image resolution of 256, for both evaluation and optimization. In the main paper, all LPIPS scores have been multiplied by 100 for readibility.

(C)SFID. The FID metric [23] measures the distance between the distributions of the real images and generated images in the feature space of an InceptionV3 classifier [47].

More formally, let μ^r and σ^r be the mean and standard deviation of inception features for the real images, and μ^s and σ^s for the synthetic images. The Simplified FID [33] is computed as

$$SFID(\alpha) = \|\mu^r - \mu^s\|^2 + \alpha\|\sigma^r - \sigma^s\|^2. \quad (6)$$

It does not take into account the off-diagonal terms in the feature covariance matrix to avoid numerical instability.

The Conditional Simplified FID (CSFID) is computed in the same manner but for each target class separately, and then averaging the resulting scores: With μ_c^r and σ_c^r the mean and standard deviation of inception features for the real images belonging to class c, and μ_c^s and σ_c^s for the synthetic images, we have

$$CSFID(\alpha) = \frac{1}{|C|}\sum_c \|\mu_c^r - \mu_c^s\|^2 + \frac{\alpha}{|C|}\sum_c \|\sigma_c^r - \sigma_c^s\|^2. \quad (7)$$

We have noticed that the distance on standard deviations was not very discriminative: since we are modifying images and not generating images from scratch, we already have a lot of diversity in the generated images. Experimentally, using $\alpha > 0$ mostly consisted in adding a bias term in this metric, therefore we chose to use $\alpha = 0$ in the (C)SFID scores.

Since the images we transform are extracted from the ImageNet validation set, we use the ImageNet training set as our reference distribution to compute the (C)SFID scores. As for LPIPS, the (C)SFID scores are computed at an image resolution of 256.

Accuracy. We use a DeiT classifier [48] trained on ImageNet, which takes images of size 384×384. Smaller images are upsampled before being passed to the classifier.

A.4. Details on the multimodal encoder

For data augmentations, we use a random horizontal flipping and a random rotation between -10 and 10 degrees, followed by cropping the image (keeping at least 80% of the input image) with aspect ratio between 0.9 and 1.1. For the CLIP-based multimodal encoders, we have considered all CLIP networks freely available, listed in Table 4.

Backbone	Params.	Latent dim.
RN50	38M	512
RN50x4	87M	640
ViT-B/32	88M	512
ViT-B/16	86M	512
RN50x16	167M	768

Table 4. Visual backbones used for the multimodal encoder. Our default configuration only includes the ViT-B/32, the RN50 and the RN50x4.

A.5. Additional qualitative results

In Figure 12, we show qualitative results when we replace the VQGAN image encoder with other GAN-based encoders. VQGAN has a native encoder and decoder, and thus the initial latent vector is obtained directly. For StyleGAN2 [31], we use the e4e encoder [49] followed by an additional 1,000 steps of LPIPS minimization. For the IC-GAN [6] model, we use the BigGAN [4] backbone as generator. IC-GAN is naturally conditioned on the SwaV embedding [5] of the input image; for added robustness we sample 1,000 latent points and choose the one yielding smallest LPIPS distance with respect to the input image.

Figure 13 shows intermediate transformation results with FlexIT for 0, 8, 16, 32 and 160 optimization steps. The result after zero optimization steps shows the effect of autoencoding the input image, without changing the latent representation. Figure 14 show representative failure cases for our method, due to either the regularization method or the multimodal embedding space.

Asset Name	Link
ImageNet	https://www.image-net.org
Cars	https://ai.stanford.edu/ jkrause/cars/car_dataset.html
LPIPS	https://github.com/richzhang/PerceptualSimilarity
FID	https://github.com/mseitzer/pytorch-fid
DeiT	https://github.com/facebookresearch/deit
CLIP	https://github.com/openai/CLIP
VQGAN	https://github.com/CompVis/taming-transformers
IC-GAN	https://github.com/facebookresearch/ic_gan
StyleGAN2	https://github.com/justinpinkney/awesome-pretrained-stylegan2
e4e	https://github.com/omertov/encoder4editing

Table 2. List of asset links.

Asset Name	Asset type	License
ImageNet	Images	https://www.image-net.org/download.php
Cars	Images	https://ai.stanford.edu/ jkrause/cars/car_dataset.html
LPIPS	Code and Models	BSD-2-Clause License
FID	Code and Models	Apache-2.0 License
DeiT	Code and Models	Apache License 2.0
CLIP	Code and Models	MIT License
VQGAN	Code and Models	MIT License
IC-GAN	Code and Models	Attribution-NonCommercial 4.0 International
StyleGAN2	Code and Models	https://github.com/justinpinkney/awesome-pretrained-stylegan2
e4e	Code	MIT License

Table 3. List of asset licenses.

A.6. Ablation results

In Table 5, we show ablation experiments for all FlexIT parameters; this includes the CSFID scores of the hyperparameter configurations reported in Figure 11 of the main paper.

In Table 6, we show ablations for combining multiple CLIP networks and using multiple data augmentations in the multimodal encoder. This includes the CSFID scores reported in Figure 9 of the main paper; we also report the runtime needed for each algorithm.

A.7. Further evaluation on COCO

We argue in section 2 that evaluating text-driven image editing is a difficult task that requires (1) a list of sensible transformation queries, and (2) a method for evaluating the quality and accuracy of the generated result. We have found another evaluation protocol in the ManiGAN paper [39], using the COCO dataset, that consists in (1) choosing *random* COCO captions/image pairs and thus leading to noisy transformations and (2) calculating the image-text similarity score which was used as a loss term during their training, leading to bias in the final scores. In the main paper, we compare different methods using our novel evaluation protocol, which was carefully designed to avoid these pitfalls. Nonetheless, we show in Tab. 7 that even with this COCO evaluation, FlexIT improves upon the scores of ManiGAN by a large margin.

	IS↑	SIM↑	DIFF↓	MP↑
ManiGAN	14.96	0.087	0.216	0.068
FlexIT	**18.19**	**0.177**	**0.146**	**0.151**

Table 7. ManiGAN evaluation on random transformation requests from COCO.

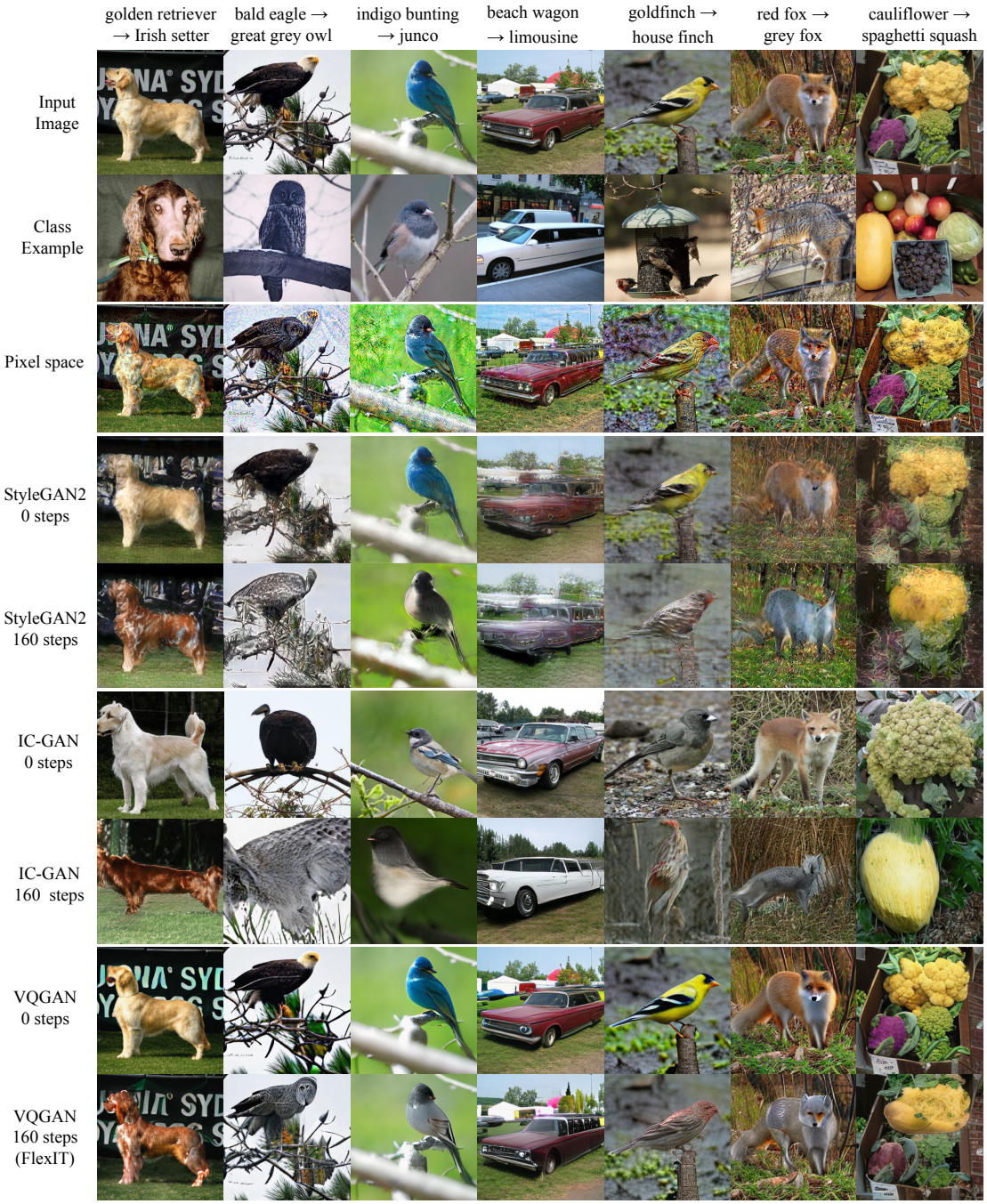

Figure 12. Transformation examples with various backbones for the image latent space. For each latent space, we show the initial image decoded from the initial point z_0, and the resulting image after 160 optimization steps. The three latent spaces differ substantially in their encoding images (0 steps). The IC-GAN latent space provides natural images that are far away from the input image due to the limited generator capacity in conjunction with the smaller latent space size (2560 dim.). StyleGAN2 images preserve the input image appearance thanks to the larger size of its latent space $\mathcal{W}+$ (8192), however images contain many unnatural artifacts due to the challenges of embedding images in this latent space [49]. The VQGAN latent space leads to the best reconstruction results. After 160 steps of optimization, the images generated with StyleGAN2 still have the same unnatural artifacts, and images generated with IC-GAN remain natural but far from the input images. VQGAN, which we use in FlexIT, achieves good edits while preserving the overall image appearance. The pixel-space method introduces high-frequency artifacts, without substantially modifying the high-level semantic image content, resembling adversarial examples for image classification.

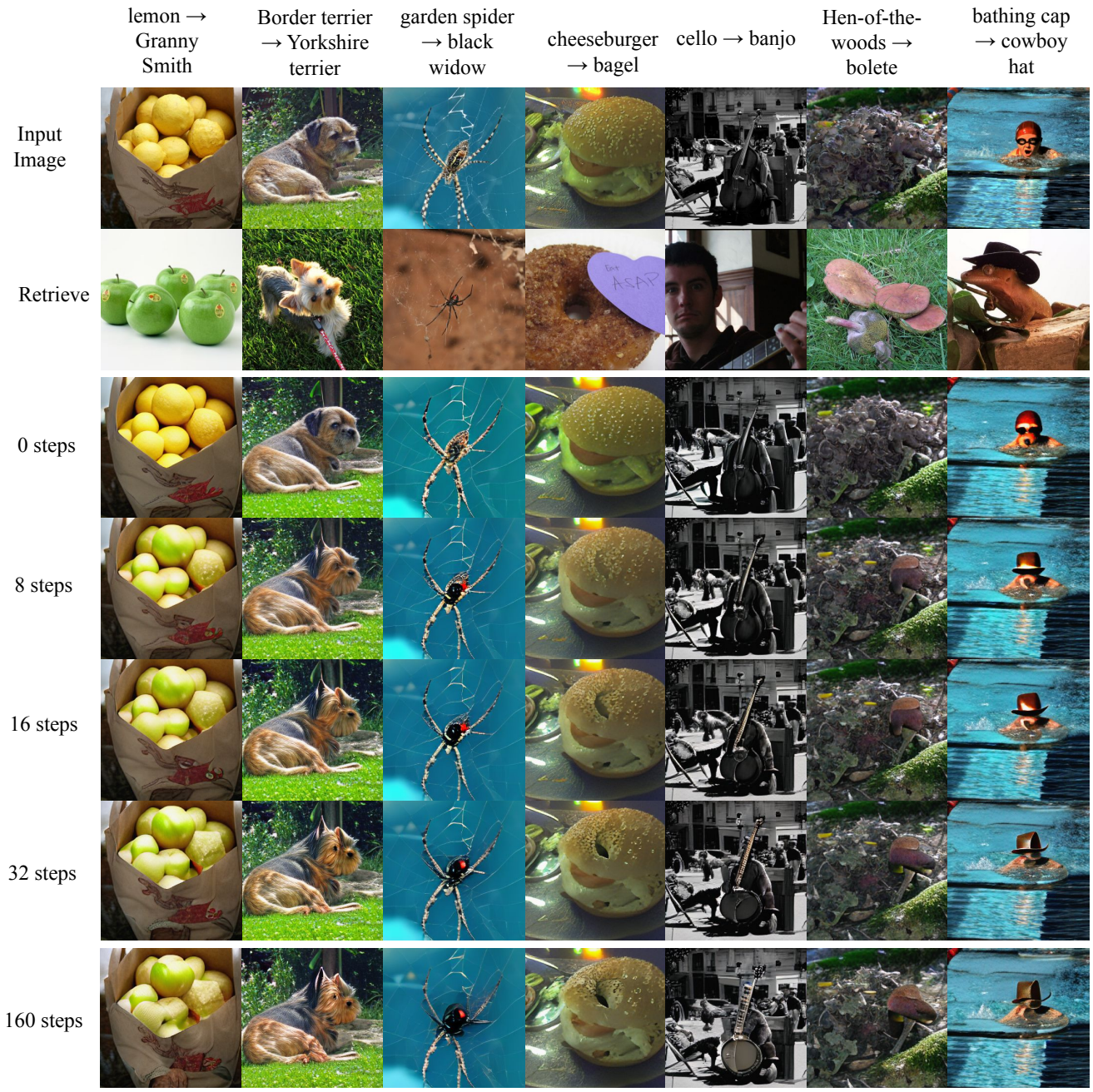

Figure 13. Intermediate transformation results obtained with FlexIT. Note that most edits only require 32 steps to be completed; some edits benefit from longer optimization schemes, such as the spider and the banjo.

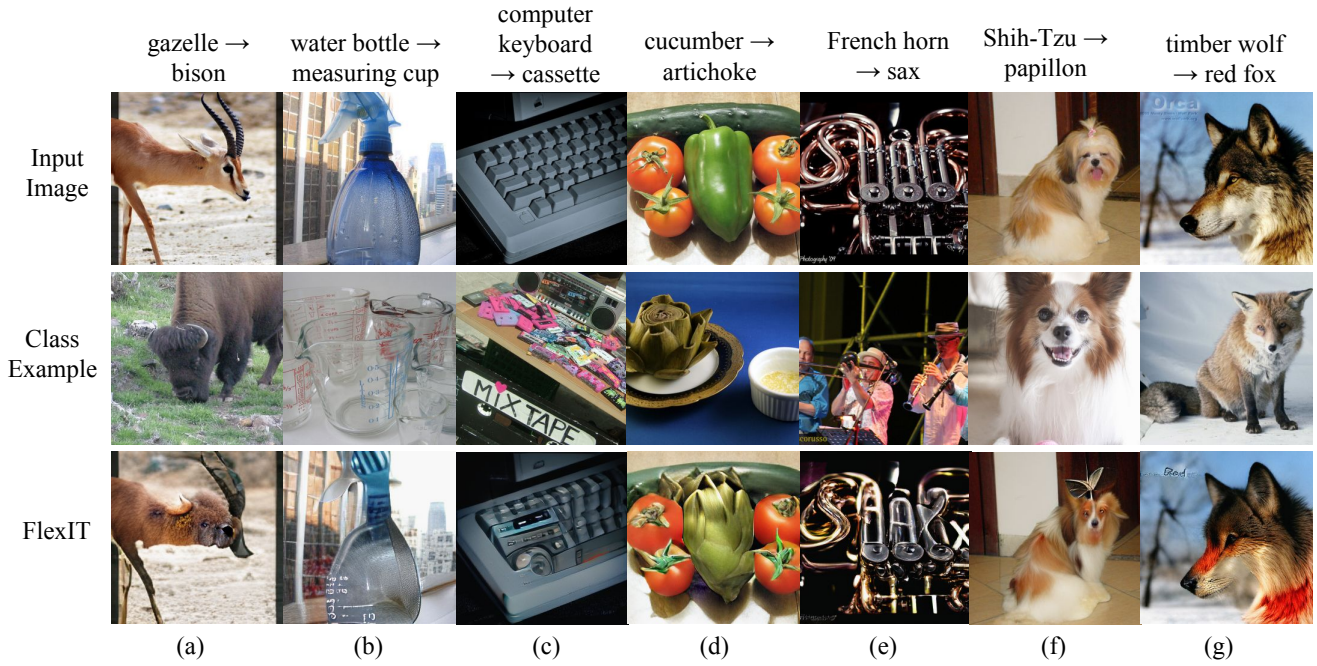

Figure 14. Representative failure cases of FlexIT. The first three columns show examples where the regularization with respect to the initial image was too strong. (a): FlexIT added bison-like texture but fails to change the shape convincingly. (b): markings have been added to the bottle, but without changing its shape to that of a measuring cup. (c): only a part of the input object was changed. (d): the bell pepper rather than the cucumber was transformed, probably because the former is more centered, and has a better initial shape. Columns (e)–(g) show failure cases related to the CLIP embedding space. (e): we observe an interesting text synthesis behaviour where the letters of the target class "sax" have been written in the image. This is related to the OCR capabilities of CLIP. (f): a butterfly is synthesized on the head of the dog (CLIP optimized for both the dog breed papillon and the insect papillon). (g): an unrealistic image is produced by adding saturated red to the image.

	Acc.↑	LPIPS↓	CSFID↓	SFID↓
$\lambda_I = 0$	**64.8**	27.6	65.4	12.3
$\lambda_I = 0.1$	60.6	25.9	57.8	8.3
$\lambda_I = 0.2$	52.6	24.6	**55.9**	6.4
$\lambda_I = 0.3$	45.8	23.5	56.3	5.5
$\lambda_I = 0.4$	38.6	**22.6**	58.6	**5.0**
$\lambda_S = 0.0$	34.3	**23.8**	60.2	**4.8**
$\lambda_S = 0.2$	45.9	24.0	57.3	5.5
$\lambda_S = 0.4$	52.6	24.6	**55.9**	6.4
$\lambda_S = 0.5$	56.2	25.0	56.5	7.1
$\lambda_S = 0.8$	**60.0**	26.5	65.5	11.7
$\lambda_z = 0.0$	**59.4**	26.5	56.1	7.1
$\lambda_z = 0.05$	52.6	24.6	**55.9**	6.4
$\lambda_z = 0.1$	51.0	**23.3**	56.7	**6.3**
$\lambda_p = 0.05$	**66.2**	28.8	**56.0**	7.9
$\lambda_p = 0.1$	59.1	26.4	**56.0**	7.2
$\lambda_p = 0.15$	52.6	24.6	**55.9**	6.4
$\lambda_p = 0.2$	47.9	**23.3**	57.5	**6.3**
ℓ_1	**54.2**	**24.6**	56.3	6.5
ℓ_2	52.4	**24.5**	**55.9**	6.8
$\ell_{2,1}$	52.6	**24.6**	**55.9**	**6.4**
$lr = 0.025$	47.6	**22.5**	58.3	**6.0**
$lr = 0.5$	52.6	24.6	55.9	6.4
$lr = 0.1$	**60.4**	27.6	**54.8**	7.2
resolution 256	53.8	24.8	56.8	7.2
resolution 288	**52.6**	24.6	**55.9**	**6.4**
resolution 320	54.3	**24.0**	57.4	7.3

Table 5. FlexIT ablation results. lr is the learning rate. Lines corresponding to our default configuration are marked in light grey. The norms ℓ_1, ℓ_2, and $\ell_{2,1}$ refer to the distance used for regularization in the VQGAN latent space. Best values for each metric are shown in bold inside each group of parameter values.

networks	d	Acc.↑	LPIPS↓	CSFID↓	SFID↓	sec. /im
ViT-B/32	0	9.4	21.8	92.7	7.4	27s
ViT-B/32	1	37.5	26.4	76.5	11.1	27s
ViT-B/32	8	35.1	25.4	76.9	10.7	33s
ViT-B/32	32	35.5	25.0	77.7	10.8	53s
RN50x4	0	13.4	23.8	91.6	11.8	35s
RN50x4	1	32.5	27.4	80.2	13.7	35s
RN50x4	8	31.0	25.2	77.3	12.3	53s
RN50x4	32	27.0	24.2	79.1	11.7	122s
2 nets	0	23.0	22.8	80.7	9.5	39s
2 nets	1	50.6	26.4	63.2	8.9	39s
2 nets	8	47.8	24.9	62.7	8.4	64s
2 nets	32	47.4	24.2	62.9	8.1	160s
3 nets	0	30.4	22.5	72.2	8.3	45s
3 nets	1	54.9	26.0	56.7	6.7	45s
3 nets	8	52.6	24.6	55.9	6.4	75s
3 nets	32	51.7	24.0	56.7	6.7	190s
5 nets	0	39.6	22.4	66.8	7.7	70s
5 nets	1	60.3	25.5	51.9	5.5	70s
5 nets	8	60.1	23.9	52.1	5.4	176s
5 nets	32	52.0	22.8	52.7	5.2	560s

Table 6. Ablation results for the multimodal encoder components. d is the number of augmentations. $d = 0$ means that the encoder takes the unchanged image as input; For $d = 1$, the encoder takes only one (augmented image), which explains why the edit time is the same as $d = 0$. When considering n CLIP networks, we take the first n elements in the following list: RN50x4, ViT-B/32, RN50, ViT-B/16, RN50x16. Our default configuration is marked in light grey. Last column gives computation time per image in seconds.

Group	Cluster	Classes
bird	bird of prey	bald eagle, kite, great grey owl
bird	finch	indigo bunting, goldfinch, house finch, junco
bird	grouse	black grouse, prairie chicken, ptarmigan, ruffed grouse
bird	seabird	king penguin, albatross, pelican, European gallinule, black swan
bird	wading bird	goose, oystercatcher, little blue heron, black stork, bustard, flamingo, spoonbill
container	bag	backpack, plastic bag, purse
container	food container	water jug, beer bottle, water bottle, wine bottle, coffee mug, vase, coffeepot, teapot, measuring cup, cocktail shaker
device	electronics	cassette player, cellular telephone, computer keyboard, desktop computer, dial telephone, hard disc, iPod, laptop
device	measuring	analog clock, digital clock, wall clock, stopwatch, digital watch, odometer, barometer
dog	hound	English foxhound, Italian greyhound, Afghan hound, basset, beagle, otterhound
dog	sporting dog	English springer, cocker spaniel, golden retriever, Irish setter
dog	terrier	American Staffordshire terrier, wire-haired fox terrier, standard schnauzer, Border terrier, Irish terrier, Yorkshire terrier
dog	toy dog	papillon, Chihuahua, Japanese spaniel, Shih-Tzu, toy terrier
dog	working dog	collie, German shepherd, Rottweiler, miniature pinscher, French bulldog, Siberian husky, boxer, Eskimo dog
edible	edible fruit	Granny Smith, strawberry, lemon, orange, banana, custard apple, fig, pineapple, pomegranate
edible	sandwich	cheeseburger, hotdog, bagel
edible	vegetable	bell pepper, broccoli, cauliflower, spaghetti squash, zucchini, butternut squash, artichoke, cardoon, cucumber
fungus	fungus	bolete, coral fungus, earthstar, gyromitra, hen-of-the-woods, stinkhorn
insect	beetle	ground beetle, ladybug, leaf beetle, long-horned beetle, tiger beetle, weevil
insect	butterfly	monarch, admiral, cabbage butterfly, lycaenid, ringlet, sulphur butterfly
insect	spider	black widow, garden spider, tarantula, wolf spider, scorpion
mammal	bear	American black bear, brown bear, ice bear, sloth bear, giant panda, lesser panda
mammal	bovid	ox, ibex, bighorn, gazelle, impala, water buffalo, ram, bison
mammal	canine	Arctic fox, grey fox, red fox, African hunting dog, dingo, coyote, red wolf, timber wolf, white wolf, hyena
mammal	equine	sorrel, zebra
mammal	feline	Persian cat, tabby, cheetah, jaguar, leopard, lion, snow leopard, tiger
mammal	great ape	chimpanzee, gorilla, orangutan
mammal	monkey	capuchin, spider monkey, squirrel monkey, baboon, guenon, macaque
music. instr.	percussion	chime, drum, gong, maraca, marimba, steel drum
music. instr.	stringed	cello, violin, acoustic guitar, electric guitar, banjo
music. instr.	wind	bassoon, oboe, sax, flute, cornet, French horn, trombone
object	ball	golf ball, ping-pong ball, rugby ball, soccer ball, tennis ball
object	handtool	hammer, plane, plunger, screwdriver, shovel
object	headdress	bathing cap, shower cap, bonnet, cowboy hat, sombrero, football helmet
reptile	amphibian	bullfrog, tree frog, axolotl, spotted salamander, common newt, eft, European fire salamander
reptile	snake	rock python, boa constrictor, green mamba, Indian cobra, diamondback, sidewinder, horned viper, king snake, green snake, thunder snake
reptile	turtle	box turtle, mud turtle, terrapin
sea life	aqu. mammal	killer whale, grey whale, sea lion, dugong
sea life	bony fish	goldfish, tench, eel, anemone fish, lionfish, gar, sturgeon
sea-life	crab	American lobster, Dungeness crab, fiddler crab, king crab, rock crab, crayfish, hermit crab, isopod
sea life	shark	great white shark, tiger shark, hammerhead
vehicle	bicycle	motor scooter, tricycle, unicycle, mountain bike, moped
vehicle	boat	speedboat, lifeboat, canoe, fireboat, gondola
vehicle	car	ambulance, beach wagon, cab, convertible, jeep, limousine, minivan, sports car
vehicle	locomotive	electric locomotive, steam locomotive
vehicle	sailing vessel	catamaran, trimaran, schooner
vehicle	truck	minivan, police van, fire engine, garbage truck, pickup, tow truck, trailer truck, school bus

Table 8. Groups and clusters of the ImageNet classes used to define the transformation queries.