Acknowledgements. We extend our sincere gratitude to Yuanzhen Li for helping with computational resources. We also appreciate the insightful reviews and valuable feedback on an early draft of the paper by Mathilde Caron and Sergey Ioffe. Additionally, we are grateful for the unwavering support and collaboration offered by the CCI team at Google Research and the members of Isolab at MIT.

A. Concept Sampling

The concepts used to synthesize captions are randomly sampled from the names of various datasets. The rough ratios are presented in Table A1. It is likely that different combinations of these ratios lead to different results, but we do not optimize over this dimension. For example, we simply concatenate IN-21k concepts with the classes of *other* datasets (*e.g.*, Caltech-101, Pets), and do uniform sampling from the concatenated list. This may lead to under-sampling for *other* datasets, as the list is dominated by IN-21 classes.

source	prob.
IN-1k	0.47
Aircraft	0.05
Cars	0.05
Food	0.05
Flowers	0.03
Places-365, SUN397	0.09
IN-21k and others	0.26

Table A1. **Rough concept sampling probabilities.**

B. Implementation Details

B.1. Pre-training

The setting for our final long schedule training in Section 4.2 is summarized in Table A2, where models are trained for 500k steps with a batch size of 8192 captions. For ablation study present in Section 4.1, we only train for 85k steps with a batch size of 2048 captions; for the scaling plots in Section 4.3, we train all models for 300k steps with a batch size of 2048.

B.2. ImageNet linear probing

We use the `cls` token from the final transformer block as the image representation. This is different from DINO v2, which tries to concatenate `cls` token with average pooled patch tokens and sweep over whether to use multiple layers.

We follow prior work [3, 5] to train the linear classifier. It has been generally observed that regularization such as weight decay hurts the performance [12, 21]. Therefore, we set weight decay as 0, and we sweep the *base_lr* over $\{0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50\} \times 10^{-2}$.

config	value
batch size	8192
optimizer	AdamW [17]
peak learning rate	2e-3 (B), 1.5e-3 (L)
weight decay	0.04 –> 0.2, cosine
optimizer momentum	β_1, β_2=0.9, 0.999
learning rate schedule	cosine decay
steps	500k
warmup steps	80k
stoch. depth [15]	0.1 (B), 0.4 (L)
augmentation	Downsample [22] + BYOL Aug. [11]

Table A2. **SynCLR pre-training settings.**

config	value
batch size	1024
optimizer	SGD
base learning rate	sweep
peak learning rate	$blr \times bsz/256$
weight decay	0
optimizer momentum	0.9
learning rate schedule	cosine decay
epochs	90
augmentation	RandomResizedCrop, Flip

Table A3. **ImageNet linear probing settings.**

B.3. End-to-End ImageNet fine-tuning

Following common practice [2, 13], we append a linear classifier on top of the `CLS` token of the last transformer block, and fine-tune the whole network. We use layer-wise *lr* decay [7]. Table A4 shows the settings.

B.4. Semantic segmentation on ADE20k

We conduct the experiments on ADE20k [27]. Following [2, 13], we use UperNet [23] as the task adaptation layer. We use the common *single-scale* [2] setup, with a resolution of 512×512 for models with a patch size of 16×16 and a resolution of 518×518 for models with a patch size of 14×14. The hyper-parameters are summarized in Table A5.

B.5. Fine-grained linear classification

Following prior works [4, 11], we train a regularized multinomial logistic regression model upon the output `CLS` token. In training and testing, we do not perform any data augmentation; images are resized to 224 pixels along the shorter side, followed by a center crop of 224×224. We minimize the cross-entropy objective using L-BFGS with ℓ_2-regularization. We select this ℓ_2-regularization constant on the validation set over 45 logarithmically spaced values between 10^{-6} and 10^{5}. The maximum number of L-BFGS iterations is set to 1000, similar as that in DINO v2 [18].

config	value
optimizer	AdamW [17]
base learning rate	5e-5
peak learning rate	$blr \times bsz/256$
optimizer momentum	$\beta_1, \beta_2{=}0.9, 0.999$
layer-wise lr decay	0.65 (B), 0.8 (L)
batch size	1024
learning rate schedule	cosine decay
warmup epochs	20 (B), 5 (L)
epochs	100 (B), 50 (L)
RandAugment [8]	9/0.5
label smoothing	0.1 (B), 0.2 (L)
erasing prob.	0.25
mixup [26]	0.8
cutmix [25]	1.0
stoch. depth [15]	0.1 (B), 0.3 (L)
test crop ratio	0.95 (B), 1.0 (L)
ema	0.9999

Table A4. **ImageNet end-to-end fine-tuning settings.**

config	value
batch size	32 (B), 16 (L)
optimizer	AdamW [17]
peak learning rate	8e-5
optimizer momentum	$\beta_1, \beta_2{=}0.9, 0.999$
weight decay	0.05
layer-wise lr decay	0.6 (B), 0.8 (L)
steps	60k (B), 160k (L)
warmup steps	1500
stoch. depth	0.1 (B), 0.2 (L)

Table A5. **ADE20k semantic segmentation settings.**

method	pre-train data	ViT-B	ViT-L
MoCo v3	real, IN1K-1M	83.2	84.1
SimMIM	real, IN1K-1M	83.8	-
MAE	real, IN1K-1M	83.6	85.9
PeCo	real, IN1K-1M	83.6	85.9
data2vec	real, IN1K-1M	84.2	86.6
iBOT	real, IN21K-14M	84.4	86.6
BEiT v2	real, WIT-400M+IN1k-1M	85.5	87.3
CLIP	real, WIT-400M	85.2	$87.5^{\dagger}$
OpenCLIP	real, LAION-400M	85.0	$86.6^{\dagger}$
	real, LAION-2B	-	$87.1^{\dagger}$
SynCLR	synthetic, 600M	**85.8**	**$87.9^{\dagger}$**

Table A6. **Top-1 accuracy on ImageNet with fine-tuning evaluation.**. Models are fine-tuned at 224x224 resolution. † use patch size of 14x14.

C. ImageNet Fine-tuning.

		EuroSAT	GTSRB	Country211	MNIST	RESISC45	KITTI	Average
CLIP	ViT-B/16	97.1	86.6	33.3	99.0	92.7	64.7	**78.9**
	ViT-L/14	98.2	92.5	42.9	99.2	94.1	69.2	**82.7**
DINO v2	ViT-B/14	96.0	72.8	21.6	98.6	92.5	75.3	76.1
	ViT-L/14	96.7	74.1	24.1	98.2	93.8	76.9	77.3
SynCLR	ViT-B/16	96.6	78.6	21.0	98.4	93.7	77.3	77.6
	ViT-L/14	96.7	79.2	24.3	98.5	93.8	78.0	78.4

Table A7. **Generalization to concepts not seen by DINO v2 and SynCLR.** SynCLR outperforms DINO v2. CLIP achieves the best accuracy, possibly because its training data includes similar concepts as these datasets.

	SynCLR		CLIP	
	IN	avg.	IN	avg.
SynCaps-150M	**80.7**	**89.0**	78.3	87.7
Laion-400M	78.9	86.5	76.6	84.9

Table A8. **Compare SynCLR with CLIP on the same synthetic data.** We observe that: (1) SynCLR outperforms CLIP; (2) in our setup, *i.e.*, generating 4 images per caption, SynCaps-150M yields better representations for both SynCLR and CLIP.

Here we show the detailed ImageNet fine-tuning performance comaprison between SynCLR and other state of the art self-supervised methods [1, 2, 5, 9, 13, 24, 28] in Table A6. Our model has better performance than models trained on ImageNet images or other large scale image datasets.

D. Further analysis

SynCLR requires a list of concepts C to start off. But how will SynCLR transfer to concepts outside our list?

Generalize to unseen concepts. We consider additional datasets whose classes are outside the synthesis list, including EuroSAT [14], GTSRB [20], Country211 [19], MNIST [16], RESISC45 [6], and KITTI distances [10]. These datasets, except for KITTI, are also outside the curation list of DINO v2. Therefore, it is also a generalization test for DINO v2. Table A7 shows the linear probing results. SynCLR outperforms DINO v2 by 1.5% for ViT-B and 1.1% for ViT-L, respectively. This suggests the representations of SynCLR generalize. CLIP outperforms SynCLR and DINO v2, with most gains coming from Country211. An explanation is CLIP's training data contains similar country flags which are not in the training sets of SynCLR and DINO v2.

Given that both captions and images are synthesized, a natural question arises: how would CLIP training perform on such data?

Compare to CLIP training. We use the same data to train a ViT-B CLIP model. For each caption, we randomly choose 1 out of the 4 synthesized images in each iteration. Following common practice [19], we train for 32 epochs with a

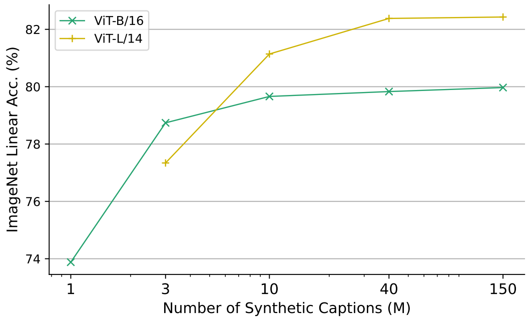

Figure A1. **ImageNet linear** accuracy w/ different training scales.

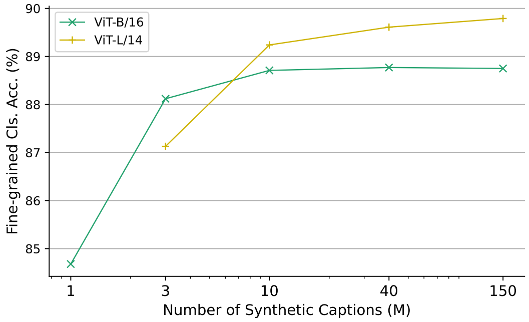

Figure A2. **Fine-grained classification** w/ different training scales.

batch size of 32768. This model achieves 44.4% zero-shot accuracy on IN-1k. The *SynCaps-150M* row in Table A8 presents the linear probing results. Synthetic CLIP learns reasonably good features, reaching 78.3% on IN-1k and 87.7% on fine-grained datasets. However, SynCLR is still better.

We have also repeated our experiments with Laion-400M captions, *i.e*., generate 4 images for each caption and train SynCLR and CLIP. The comparison between rows *SynCaps-150M* and *Laion-400M* in Table A8 suggests synthetic captions are also favorable on a large scale.

Scaling behavior. We train ViT-BViT-L models using random subsets of varying sizes: 1M, 3M, 10M, 40M, and the comprehensive 150M (measured in the number of captions). These models are trained over a reduced schedule of 300,000 steps and utilizes a smaller batch size of 2048. The outcomes of linear probing are illustrated in Figures A1 and A2. These results indicate that the ViT-B model delivers robust performance at the 10M scale, with diminishing returns observed beyond this point. In contrast, the ViT-L model exhibits a greater demand for data (i.e., it underperforms ViT-B at the 3M scale) and scales better with data.

E. In-context Learning Examples

All of the three types of in-context examples are summarized in Table A9, Table A10, and Table A11, respectively.

Table A9. Detailed in-context learning examples for Template 1: *c* –> *Caption*. Here *c* is the concept.

1	coucal	–>	A vibrant coucal is perched on the branch of a lush green tree, surrounded by wildflowers.
2	bee eater	–>	A lively bee eater is elegantly perched on a branch, peering intently.
3	three-toed sloth	–>	A three-toed sloth is lazily hanging from a sturdy, tropical rainforest tree.
4	hay	–>	In the serene countryside, hundreds of neatly stacked hay bales lay scattered under the softly glowing golden sunset sky.
5	station wagon	–>	A shiny, red station wagon is parked under the dappled shade of a large oak tree, highlighting its spacious and family-friendly design.
6	zebra	–>	A zebra is gallantly trotting across the vast, sunlit plains of the African savannah, creating a captivating black and white spectacle.
7	vase	–>	In the well-lit living room, a beautifully designed, delicate vase stands out as the centerpiece, exuding an aura of elegance.
8	barber chair	–>	A shiny black barber chair sits invitingly in a bustling, well-lit barbershop.
9	carbonara	–>	A heaping plate of creamy carbonara pasta topped with fresh parsley sprigs.
10	mink	–>	In the midst of a dense forest with shimmering green leaves, a sleek mink gracefully navigates the underbrush, showcasing its rich, brown fur.
11	small white butterfly	–>	A small white butterfly gracefully flutters amongst vibrant, blooming summer flowers.
12	christmas stocking	–>	A vibrant red Christmas stocking is hanging delicately from a festively decorated mantelpiece.
13	horse-drawn vehicle	–>	An antique horse-drawn vehicle is stationed amidst a peaceful country landscape, its rustic wooden structure gleaming under the warm afternoon sun.
14	ruler measuring stick	–>	A manual craftsman is precisely measuring a wooden log with a ruler stick.
15	picket fence	–>	A tranquil suburban scene featuring multiple white picket fences surrounding well-maintained green lawns, punctuated by diverse, colorful flowerbeds.
16	suspension bridge	–>	Depicting a long suspension bridge, its steel cables elegantly stretching towards the sky, connecting two ends over a scenic river.
17	brain coral	–>	A vibrant brain coral stands out amidst the serene backdrop of underwater marine life.
18	revolver	–>	Multiple antique revolvers lie on a wooden table, gleaming under soft, ambient light.
19	slip-on shoe	–>	A pair of slip-on shoes, with their sleek, black leather exterior and comfortable, cushioned interior, are neatly placed on a wooden floor.
20	hand-held computer	–>	A hand-held computer, compact and portable, rests on a well-lit desk, surrounded by various technological paraphernalia and a steaming cup of coffee.
21	mattress	–>	A teddy bear lying face down on a bedspread covered mattress in front of a window.
22	refrigerator	–>	A nicely decorated kitchen with metallic refrigerator and blue counter.
23	ball	–>	Silver balls are lined up in the sand as people mill about in the background.
24	wheel	–>	The motorcycle's gleaming steering wheel, vivid red door reflected in the side mirror, and a youth passing by, creating a dynamic urban tableau.
25	plane	–>	A group of trick planes turned upside down leaving smoke trails.
26	vehicle	–>	Army vehicles, including a U.S. Army jeep and aircraft in a hangar or on display
27	boy	–>	a little boy wearing sunglasses laying on a shelf in a basement.
28	fence	–>	a man standing near a fence as reflected in a side-view mirror of a red car.
29	wood table	–>	A footed glass with water in front of a glass with ice tea, and green serpentine bottle with pink flowers, all on a wood table in front of chair, with a window to city view.
30	toilet	–>	A black and white toilet sitting in a bathroom next to a plant filled with waste.
31	table lamp	–>	A textured brass table lamp, casting a warm, golden glow, accents a cozy reading nook beside a leather armchair and a stack of books.
32	hair dryer	–>	A modern sleek and white hair dryer, with a textured grip, stands next to a set of hairbrushes.
33	street sign	–>	The street signs indicate which way a car can and cannot turn while the signal light controls traffic.
34	instrument	–>	Man dressed in Native American clothes protecting musical instruments from the rain with an umbrella.

35 train –> A man and a cow's faces are near each other as a train passes by on a bridge.

36 giraffe –> A couple of large giraffe standing next to each other.

37 red admiral butterfly –> a red admiral butterfly, alights upon a dew-kissed sunflower, wings glistening under the soft morning light.

38 stupa –> Surrounded by verdant foliage, a white stupa rises, adorned with golden accents and intricate patterns, while devotees circle its base offering prayers.

39 elephant –> A group of elephants being led into the water.

40 bottle –> Motorcycles parked on a street with a bottle sitting on the seat of the nearest the camera.

41 trombone –> On a polished wooden stage, a gleaming brass trombone rests, its slide extended, next to scattered sheet music and a muted trumpet.

42 keyboard –> Sleek black keyboard with illuminated backlit keys, a soft wrist rest, and a nearby wireless mouse on a textured matte desk surface.

43 bear –> The brown bear sits watching another bear climb the rocks

44 snowboard –> A man standing next to his snowboard posing for the camera.

45 railway –> a woman and her son walking along the tracks of a disused railway.

46 sand –> the waves and the sand on the beach close up

47 pixel –> very colorful series of squares or pixels in all the colors of the spectrum , from light to dark

48 cigar –> a burning cigar in a glass ashtray with a blurred background.

49 music –> happy girl listening music on headphones and using tablet in the outdoor cafe.

50 earring –> this gorgeous pair of earrings were featured in april issue.

51 cliff –> Steep cliff, jagged edges against azure sky, with seabirds soaring and waves crashing below.

52 corn cob –> Fresh corn cob, golden kernels glistening with dew, nestled amid green husks in a sunlit field.

53 archaeological excavation –> In this intriguing scene, archaeologists meticulously uncover ancient relics at an archaeological excavation site filled with historical secrets and enigmas.

54 formal garden –> This is an immaculately kept formal garden, with perfectly trimmed hedges, colorful, well-arranged flower beds, and classic statuary, giving a vibe of tranquil sophistication.

55 veterinarians office –> The busy veterinarian's office is a hive of activity with pets awaiting treatment and care.

56 elevator –> A modern, well-lit elevator interior with shiny metal walls and sleek buttons.

57 heliport –> Situated in a lively area, the heliport stands out with numerous helicopters taking off and landing against the city's skyline.

58 airport terminal –> In the spacious airport terminal, travelers hurriedly navigate through check-ins and security, making it a hive of constant activity.

59 car interior –> Inside the car, the leather seats exude luxury, contrasted by the high-tech dashboard, creating an atmosphere of sleek comfort and convenience.

60 train interior –> The inside of the train offers a spacious setting with numerous comfortable seats.

61 candy store –> The sweet aroma of sugared treats fills the air in a vibrant candy store, adorned with colourful candies and cheerful customers.

62 bus station –> The bustling bus station thrums with restless energy, as travelers navigate through the crowded space, awaiting their journeys amid the echoes of departing buses.

63 castle –> Nestled amidst towering mountains, the majestic castle spews ancient grandeur, with its stone walls and towering turrets exuding tranquility and timeless mystique.

64 palace –> The grand palace exudes regality, radiant under the sun, showcasing ornate decorations, intricate sculptures, and exquisite architectural sophistication.

65 kitchen –> The heart of the home unfolds in the kitchen, characterized by stainless steel appliances, navy blue cabinets, and a patterned tile backsplash.

66 raceway –> The high-speed adrenaline-filled atmosphere of the raceway is pulsing with the roars of powerful engines and excited cheering fans.

67 bakery –> The warm, inviting bakery is filled with the intoxicating aroma of fresh bread, assorted pastries, and brewing coffee.

68	medina	–>	This ancient, labyrinth-like medina exudes an air of mystique with its vibrantly decorated shops lining narrow, stone-cobbled pathways.
69	skyscraper	–>	The city skyline is dominated by towering skyscrapers, creating a captivating blend of technology and architectural innovation.
70	supermarket	–>	The supermarket scene is lively, filled with individuals scanning shelves, children reaching for treats, and clerks restocking fresh produce.
71	closet	–>	The compact closet, brimming with clothes and shoes, exudes a feeling of organization.
72	assembly line	–>	In the heart of a busy factory, an orderly assembly line hums with continuous activity, filled with workers focused on their precision tasks.
73	palace room	–>	A man in military dress uniform stands in an ornate palace room with antique furniture and Christmas decorations.
74	barn doorway	–>	A farmer holding an animal back while another farmer stands in a barn doorway.
75	food court	–>	A bustling food court with a variety of culinary stalls, featuring vibrant signage, aromatic dishes, and communal seating, creates a diverse dining experience.
76	mountain	–>	Majestic mountains, their peaks dusted with snow, overlook a serene alpine lake where hikers and photographers gather to enjoy the breathtaking scenery.
77	squash court	–>	Against a clear glass wall, a squash court with gleaming wooden floors, white boundary lines, and two rackets awaits players.
78	subway station	–>	Dimly lit subway station with graffiti-covered walls, commuters waiting
79	restaurant	–>	Cozy restaurant with wooden tables, ambient lighting, patrons chatting, and plates filled with colorful dishes, framed by exposed brick walls and hanging green plants.
80	field	–>	there is a large heard of cows and a man standing on a field.
81	aquarium	–>	Amidst vivid coral formations, an aquarium teems with colorful fish, shimmering under soft blue lights.
82	market	–>	A large group of bananas on a table outside in the market.
83	park	–>	a young boy is skating on ramps at a park
84	beach	–>	old fishing boats beached on a coastal beach in countryside.
85	grass	–>	little boy sitting on the grass with drone and remote controller.
86	woven	–>	The woven basket's intricate pattern creates a visually captivating and tactile surface.
87	knitted	–>	The knitted blanket envelops with cozy warmth
88	flecked	–>	The stone surface was flecked, giving it a uniquely speckled and rough appearance.
89	bubbly	–>	The liquid gleamed, showcasing its bubbly, effervescent texture vividly.
90	cobwebbed	–>	The dusty corner was cobwebbed, displaying years of untouched, eerie beauty.
91	stained	–>	A weather-worn wall manifests an intriguing pattern of stained texture.
92	scaly	–>	The image showcases a close-up of a lizard's scaly, rough texture.
93	meshed	–>	A patterned image depicting the intricate, tightly-knit texture of meshed fabric.
94	waffled	–>	A fresh, golden-brown waffle displays its distinct crisply waffled texture invitingly.
95	pitted	–>	The image portrays an intriguing terrain, characterized by a pitted, moon-like surface.
96	studded	–>	A studded leather jacket gleams, highlighting its rough, tactile texture.
97	crystalline	–>	The picture showcases an exquisite, crystalline texture with stunning brilliance and clarity.
98	gauzy	–>	A delicate veil of gauzy texture enhances the ethereal, dreamy atmosphere.
99	zigzagged	–>	The photo captures the zigzagged texture, emphasizing the rhythmic, sharp-edged patterns.
100	pleated	–>	A flowing skirt delicately showcasing the intricate detail of pleated texture.
101	veined	–>	A detailed image showcasing the intricate, veined texture of a leaf.
102	spiralled	–>	The spiralled texture of the seashell creates a captivating, tactile pattern.
103	lacelike	–>	The delicate veil features an intricate, lacelike texture, exuding elegant sophistication.
104	smeared	–>	A wall coated with thick, smeared paint exudes a rough texture.
105	crosshatched	–>	A worn, vintage book cover, richly crosshatched, exuding old-world charm.
106	particle	–>	abstract background of a heart made up of particles.

Table A10. Detailed in-context learning examples for Template 2: *c*,*bg* –> *caption*. Here *c* is the concept, and *bg* is the background.

107	stick insect, undergrowth	–>	A stick insect, masterfully camouflaged, clings to a fern amidst the sprawling, dense undergrowth of a lush, tropical forest.
108	black swan, public garden	–>	In the peaceful ambiance of a lush public garden, a majestic black swan gracefully glides across a shimmering emerald-green pond.
109	st. bernard, family photo	–>	In the heartwarming family photo, a gregarious St. Bernard dog is seen joyfully nestled among his adoring human companions.
110	measuring cup, food prep area	–>	In the food prep area, multiple transparent measuring cups are neatly organized on the marble countertop.
111	can opener, hotel room	–>	A sleek, stainless steel can opener is sitting on the glossy dark-wood kitchenette counter of a modern, well-appointed hotel room.
112	small white butterfly, pond side	–>	A delicate, small white butterfly flutters gracefully above the tranquil pond side, creating a serene image amidst lush greenery.
113	hair dryer, theatre	–>	A sleek, professional hair dryer is positioned center stage amidst the dramatic velvet curtains and ornate details of a bustling theatre.
114	water bottle, airport	–>	A reusable water bottle sits on the glossy surface of a bustling airport terminal counter, amidst a backdrop of hurried travelers and departure screens.
115	leonberger, horse ranch	–>	Several Leonbergers are joyfully romping around a bustling horse ranch.
116	lighter, motorhome	–>	In the cozy, cluttered environment of a well-traveled motorhome, a sleek silver lighter holds dominion on the rustic wooden table.
117	slug, foliage	–>	A solitary, glistening slug meanders slowly amidst lush, dense green foliage, leaving a slimy trail on dewy leaves in its path.
118	ring binder, education department	–>	The ring binder, filled with important documents, sits prominently on a well-organized desk in the bustling education department.
119	weimaraner, pet store	–>	A sleek, silver-gray Weimaraner is spotted curiously sniffing around various pet supplies in a well-stocked and vibrant pet store.
120	norfolk terrier, countryside	–>	A lively Norfolk terrier joyfully bounds across a lush, green countryside, its red fur contrasting vividly with the vast open surroundings.
121	dalmatian, apple orchard	–>	A lively Dalmatian is playfully darting amongst the lush rows of a bountiful apple orchard, its spots contrasting against the ruby fruits.
122	television, mountain lodge	–>	A sleek, modern television sits prominently against the rustic, wooden walls of an inviting mountain lodge, surrounded by pine-furnished decor.
123	guillotine, horror story	–>	In the shadowy landscape of a suspenseful horror story, a grim, menacing guillotine looms ominously, exuding a petrifying sense of imminent dread.
124	hot tub, condominium	–>	A luxurious hot tub is nestled in the private balcony of a high-rise condominium, boasting spectacular cityscape views.
125	leaf beetle, plant nurseries	–>	A vibrant leaf beetle is diligently navigating through a lush plant nursery, its metallic sheen contrasting against the abundant green foliage.
126	carolina anole, hiking trails	–>	A small Carolina Anole lizard basks in the warm sunlight, gracefully draped over a gnarled tree root next to a bustling hiking trail.
127	girl, laboratory	–>	teenage girl and boy working in a laboratory on an experiment.
128	tiger, forest	–>	Two tigers are running together in the forest.
129	sunset, lake	–>	Golden sunset hues reflect on a calm lake, silhouetting a lone canoeist against a backdrop of fiery clouds.
130	building, mountain	–>	town of skyline over roofs of historic buildings with the mountains in the background.
131	block plane, weathered wood	–>	A block plane, its sharp blade gleaming, rests on weathered wood
132	olive tree, soil	–>	single olive tree planted in the center of a dry and cracked soil
133	hamster, pet store	–>	A curious hamster peers out, with pet store shelves stacked with supplies behind.
134	bag, factory	–>	plastic bags production line in a factory.

135	restaurant, ocean	–>	young pretty couple dining in a romantic atmosphere at restaurant on the boat with ocean on the background
136	helicopter, burning forest	–>	a helicopter flies over a portion of burning forest.
137	pipe organ, commemoration event	–>	striking pipe organ dominates with its notes resonating, while a somber commemoration event unfolds in the backdrop
138	rotisserie, wedding reception	–>	Rotisserie turning golden meats, with a bustling wedding reception, twinkling lights, and guests mingling.
139	duck, taiga	–>	A group of ducks paddle on a tranquil pond, dense taiga and towering conifers looming in the background.
140	tiger beetle, rice fields	–>	Amidst verdant rice fields, a shimmering tiger beetle perches prominently on a dew-kissed blade of grass.
141	girl, barn	–>	slow motion clip of a girl walking with her horse through a barn
142	headmaster, graduation ceremony	–>	the headmaster addresses the graduating seniors during graduation ceremonies.
143	businessperson, music festival	–>	businessperson and guest attend music festival.
144	fountain, park	–>	Water cascades from an ornate fountain, surrounded by autumn-hued trees in a serene park.
145	speedboat, water	–>	A sleek speedboat glides on shimmering waters, powered by twin high-horsepower outboard motors.
146	pipe, beach	–>	a rusty water pipe on the beach.
147	pretzel, home kitchen	–>	Golden pretzel rests on a wooden board, with a cozy home kitchen, pots and tiled backsplash, behind.
148	forklift, paper mill	–>	A forklift transports hefty paper rolls amidst the industrial bustling paper mill.
149	lotion, therapy center	–>	Blue lotion bottles lined up at a thalasso therapy center by the ocean.
150	guinea pig, sand dunes	–>	Guinea pig exploring vast golden sand dunes, with tiny footprints trailing behind.
151	groom, wedding ceremony	–>	father of groom congratulating him after the wedding ceremony.
152	fishing boat, village	–>	fishing boats moored at fishing village a suburb of capital of the state,
153	red fox, yard	–>	wild red fox sitting on a partially snow covered front yard of a house in the suburbs of a small city
154	grey wolf, woodland areas	–>	A grey wolf prowls silently, eyes alert, through dense, misty woodland areas with moss-covered trees.
155	cheetah, edges of swamplands	–>	A cheetah crouches, poised and watchful, at the lush edges of murky swamplands.
156	wine bottle, living room	–>	in the living room, a person si opening a wine bottle with corkscrew with wooden barrel

Table A11. Detailed in-context learning examples for Template 3: c,rel –> *caption*. Here c is the concept, and rel is the relation.

157	product packet / packaging, next to	–>	A vibrant product packet, adorned with colorful labels and intricate designs, is neatly placed next to an elegant crystal glass.
158	croquet ball, behind	–>	A vivid, red croquet ball rests serenely, hiding behind a worn, rustic wooden fence in a sun-kissed, lush green lawn.
159	bassoon, in front of	–>	A beautifully crafted bassoon stands elegantly in front of a backdrop of velvet curtains, ready to perform at a concert.
160	grand piano, above	–>	A gorgeous, antique chandelier is suspended above the glossy black grand piano, illuminating it with warm, opulent light.
161	bolo tie, behind	–>	A beautifully crafted bolo tie is casually hung, indicating its previous use, behind a rustic, well-polished wooden shelf.

162	waffle iron, next to	–>	A large, black waffle iron is placed next to a sparkling glass jar filled with golden maple syrup on a wooden countertop.
163	komodo dragon, below	–>	A young child grins excitedly, peering down from a secure bridge, as a colossal Komodo dragon sprawls lazily below in the wildlife park.
164	vaulted or arched ceiling, besides	–>	Besides the grand marble statue, glimpses of an intricate vaulted or arched ceiling add to the room's majestic charm.
165	gossamer-winged butterfly, next to	–>	A lovely, vibrant gossamer-winged butterfly is gently perched next to a dew-kissed red rose in an early morning garden.
166	kit fox, in front of	–>	A group of small, fluffy, golden kit foxes is playfully gathered in front of a lush, green, towering forest backdrop.
167	koala, in	–>	A cute, fuzzy koala is visibly relaxed, nestled contentedly in the crook of a towering, lush green eucalyptus tree.
168	centipede, above	–>	A vibrant green centipede is effortlessly crawling on a tree branch, positioned distinctly above a patch of untouched fern leaves.
169	mountain bike, above	–>	A mountain bike is displayed prominently above the rustic mantlepiece, showcasing its sleek design and intricate details.
170	wallaby, above	–>	A fluffy, brown wallaby is leaping high, appearing as if it is effortlessly floating above a lush, green Australian field.
171	giant panda, on	–>	A playful giant panda is perched on a sturdy tree branch, munching on fresh green bamboo amidst the tranquil forest ambiance.
172	beagle, on	–>	A pack of adorable beagles are spotted lounging on an expansive, sunbathed meadow with colorful wildflowers sprouting around them.
173	beach, on	–>	A vivid sunset is on display over a sprawling beach, casting warm hues on the waves gently lapping at the sandy shore.
174	grey whale, on	–>	A voluminous grey whale is majestically breaching, its massive body on display against the azure backdrop of the expansive ocean.
175	tractor, in front of	–>	A bright red tractor is parked in front of a rustic, weathered barn, casting long shadows under the golden afternoon sun.
176	cabbage, besides	–>	A vibrant image portrays a lush, green cabbage, glistening with dewdrops, nestled besides a rustic, wooden crate full of freshly harvested vegetables.

References

[1] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, 2022. 2

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1, 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1

[5] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 1, 2

[6] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 2017. 2

[7] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020. 1

[8] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR workshops*, 2020. 2

[9] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, Nenghai Yu, and Baining Guo. Peco: Perceptual codebook for bert pre-training of vision transformers. In *AAAI*, 2023. 2

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2

[11] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 2020. 1

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1, 2

[14] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 2

[15] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 1, 2

[16] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 2

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1, 2

[18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[19] 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 *ICML*, 2021. 2

[20] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *IJCNN*, 2011. 2

[21] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv:1906.05849*, 2019. 1

[22] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. In *NeurIPS*, 2023. 1

[23] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 1

[24] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022. 2

[25] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 2

[26] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2

[27] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019. 1

[28] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2