Appendix

In this Appendix, we first we provide the full baseline comparison (with R@1/5/10) on Flickr30K and COCO (Sec. A). Then we show the challenges of vision-language distillation (Sec. B) by transitioning the trajectory-matching pipeline from image-only to image-text retrieval. We provide analysis on distilled images (Sec. D) and lossless distillation (Sec. C). We further extend the ablation study, analyzing components of our pipeline, i.e. distilled dataset initialization (Sec E.1), encoder backbones (Sec. E.2), pretraining (Sec. E.3) and synthetic steps (Sec. E.4). Lastly, we show additional visualizations of the distilled samples, as well as the ones under different backbones (Sec. G).

A Full Details for Distilled Performance

We provide full distillation results following Section 4.2, including image-to-text and text-to-image retrieval results R@5 and R@10 with NFNet in Tab. 6.

Table 6: **Detailed Baseline comparisons of NFNet on Flickr30K (top) and COCO (bottom).** Following Tab. 1, here we report the full details of the distilled performance on Flickr30K and COCO with NFNet and BERT.

			TR					IR				
			Coreset Selection				Dist (ours)	Coreset Selection				Dist (ours)
Dataset	#pairs	Metrics	R	H	K	F		R	H	K	F	
Flickr30K	100	R@1	1.3	1.1	0.6	1.2	**9.9** ± 0.3	1.0	0.7	0.7	0.7	**4.7** ± 0.2
		R@5	5.9	4.7	5.0	4.2	**28.3** ± 0.5	4.0	2.8	3.1	2.4	**15.7** ± 0.5
		R@10	10.1	7.9	7.6	9.7	**39.1** ± 0.7	6.5	5.3	6.1	5.6	**24.6** ± 1.0
	200	R@1	2.1	2.3	2.2	1.5	**10.2** ± 0.8	1.1	1.5	1.5	1.2	**4.6** ± 0.9
		R@5	8.7	8.4	8.2	8.4	**28.7** ± 1.0	4.8	5.5	5.4	3.1	**16.0** ± 1.6
		R@10	13.2	14.4	13.5	10.2	**41.9** ± 1.9	9.2	9.3	9.9	8.4	**25.5** ± 2.6
	500	R@1	5.2	5.1	4.9	3.6	**13.3** ± 0.6	2.4	3.0	3.5	1.8	**6.6** ± 0.3
		R@5	18.3	16.4	16.4	12.3	**32.8** ± 1.8	10.5	10	10.4	9.0	**20.2** ± 1.2
		R@10	25.7	24.3	23.3	19.3	**46.8** ± 0.8	17.4	17.0	17.3	15.9	**30.0** ± 2.1
	1000	R@1	5.2	5	5.6	3.1	**13.3** ± 1.0	3.8	4.1	4.4	3.2	**7.9** ± 0.8
		R@5	15.6	14.6	16.1	14.9	**34.8** ± 1.9	11.8	12.1	12.8	9.5	**24.1** ± 1.6
		R@10	21.4	20.4	20.8	18.9	**45.9** ± 2.5	19.9	20.0	20.4	18.7	**33.8** ± 2.0
COCO	100	R@1	0.8	0.8	1.4	0.7	**2.5** ± 0.3	0.3	0.5	0.4	0.3	**1.3** ± 0.1
		R@5	3.0	2.1	3.7	2.6	**10.0** ± 0.5	1.3	1.4	1.4	1.5	**5.4** ± 0.3
		R@10	5.0	4.9	5.5	4.8	**15.7** ± 0.4	2.7	3.5	2.5	2.5	**9.5** ± 0.5
	200	R@1	1.0	1.0	1.2	1.1	**3.3** ± 0.2	0.6	0.9	0.7	0.6	**1.7** ± 0.1
		R@5	4.0	3.6	3.8	3.5	**11.9** ± 0.6	2.3	2.4	2.1	2.8	**6.5** ± 0.4
		R@10	7.2	7.7	7.5	7.0	**19.4** ± 1.2	4.4	4.1	5.8	4.9	**12.3** ± 0.8
	500	R@1	1.9	1.9	2.5	2.1	**5.0** ± 0.4	1.1	1.7	1.1	0.8	**2.5** ± 0.5
		R@5	7.5	7.8	8.7	8.2	**17.2** ± 1.3	5.0	5.3	6.3	5.8	**8.9** ± 0.7
		R@10	12.5	13.7	14.3	13.0	**26.0** ± 1.9	8.7	9.9	10.5	8.2	**15.8** ± 1.5
	1000	R@1	1.9	2.4	2.4	1.9	**6.8** ± 0.4	1.5	1.3	1.5	0.7	**3.3** ± 0.1
		R@5	7.6	9.0	9.0	7.7	**21.9** ± 1.2	5.6	5.7	7.1	4.6	**11.9** ± 0.5
		R@10	12.7	14.0	14.1	13.0	**31.0** ± 1.5	9.6	10.1	10.9	8.0	**22.1** ± 0.9

B CIFAR10 Classification vs Retrieval Distillation

Prior work has shown remarkable distillation results on CIFAR10 (Krizhevsky et al., 2009) classification. To move from distilling image-only datasets to vision-language datasets, we first check if our method has potential in simple settings. Concretely, we convert CIFAR10 labels to captions that pair with their corresponding images. Under this formulation, the objective of classification is equivalent to that of image-to-text retrieval (TR): finding the best text given an image.

In Tab. 7, we compare CIFAR10 distillation performance for dataset size of 1, 10, 50 images per class (IPC), under three different settings: classification, single-caption retrieval, and multi-caption retrieval. For classification, we demonstrate results from MTT (Cazenavette et al., 2022), where they distill an image-only dataset using expert trajectories trained on image-label pairs. In single-caption TR, we distill image-caption pairs using expert trajectories trained when each image is paired with a single caption `"This is a {label}"`. In multi-caption TR, we distill image-caption pairs but the expert trajectories are trained when each image is paired with five captions that are generated with varies prompts from (Radford et al., 2021). For consistency, all image trajectories are obtained with the 3-layer ConvNet backbone as specified in (Cazenavette et al., 2022), and text trajectories are from linear projection layers over pretrained BERT (Devlin et al., 2018) embeddings. Although the performance of vision-language distillation trails behind that of image-only distillation, the gap closes at larger IPCs. However, this gap highlights the challenge of the continuous label space in vision-language datasets. Moreover, the performance gap between single and multi-caption retrieval demonstrates the challenge of capturing the variability within human language descriptions.

Table 7: **CIFAR10 Classification vs Retrieval.** We provided ipc=1/10/50 classification performance vs. image-to-text retrieval R@1, which both measure whether an image has been matched with the correct class.

IPC	Classification	image-to-text retrieval	
		Single Caption	Multi Caption
1	46.3 ± 0.8	27.4 ± 1.0	22.3 ± 1.0
10	65.3 ± 0.7	35.9 ± 0.7	33.2 ± 0.5
50	71.6 ± 0.2	66.8 ± 1.1	62.0 ± 0.8
Full	84.8 ± 0.1	79.6 ± 0.6	80.3 ± 0.4

C Upper Bound Performance

We further increase the distilled size to be 10% of the original Flickr30K dataset size and we provide the comparisons for distillation performance with the upper bound results (Tab. 8). The distillation performance are closely approaching the upper bound results.

Table 8: **Matching Upper Bound Performance.** With only 10% of the original size of Flickr30K, models trained on this distilled data show remarkable performance, closely approaching the upper bound results. For certain metrics, such as NFNet + CLIP with Text Retrieval (TR) at R@1, they reach as high as 98% of the upper bound.

Result	Vision	Language	Ratio	TR			IR		
Type	Backbone	Backbone		R@1	R@5	R@10	R@1	R@5	R@10
Distillation	NFNet	BERT	10%	32.1	60.0	73.2	24.1	53.9	66.5
Upper Bound	NFNet	BERT	100%	33.9	65.1	75.2	27.3	57.2	69.7
Distillation	NFNet	CLIP	10%	60.0	86.3	91.4	47.4	78.2	86.5
Upper Bound	NFNet	CLIP	100%	61.2	87.5	92.8	49.8	79.8	88.3

D Analysis on Distilled Images

We have found that increasing the learning rate and distillation time lead to more noticeable changes in the images within the distilled dataset (distilled images: Fig. 4, original images: Fig. 5). However, it is important to note that a higher learning rate or longer distillation time does not necessarily translate to improved performance of the distilled dataset, even if the images appear to deviate more drastically from the human perception perspective. Changes in image pixels alone may not reliably predict distillation performance. It is rather a measurement of the distillation strength. More distorted images suggest uneven pixel updates, while even updates yield results similar to the visualization we provided before in Fig. 3.

In line with previous studies, we initially expected more obvious changes in images would lead to better performance, but our findings suggest a different behavior of vision-language distillation with trajectory matching framework, reflecting how models capture vision-language interaction. From a human perception perspective, the distilled images appear to be moving less compared to previous classification works, yet those small vectors are still meaningful and contain useful information, as opposed to artifacts like noisy patterns. Our algorithm achieves a clear and consistent improvement over random baselines indicated by the results. We hope this discussion can inspire more research on vision-language dataset distillation.

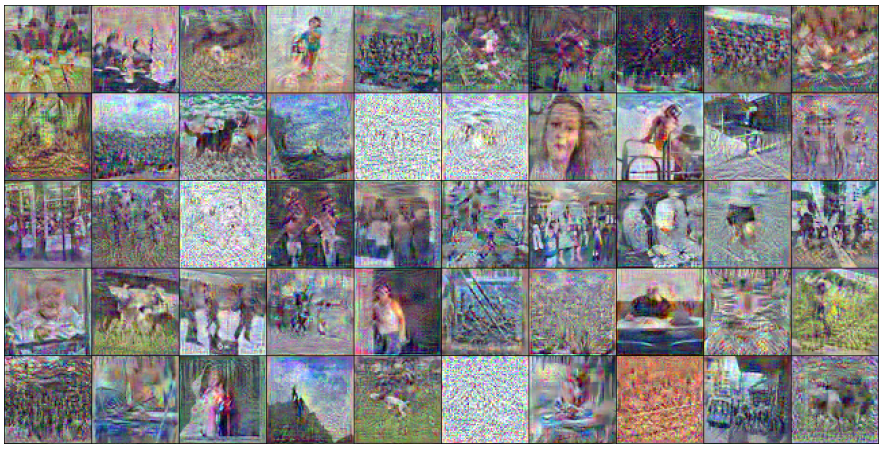

Figure 4: Distilled Images, iteration = 7000, lr image = 5000.

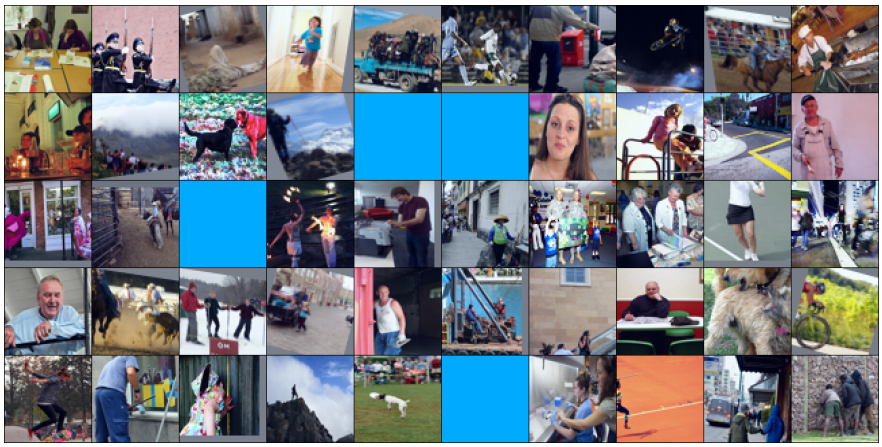

Figure 5: Original Images, iteration = 0.

E Additional Ablation Studies

In this section, we provide additional ablation studies. Unless specified, these distillation experiments are conducted on the Flickr30K dataset to distill 100 image-text pairs, and we use pretrained NFNet and BERT as backbones, with synthetic step set to 8 during distillation.

E.1 Distilled Dataset Initialization

In the main paper, we provided experiments with real sample initialization. Here we experiment and evaluate initializing with Gaussian noise. Our findings in Tab. 9 show that initializing images from the Gaussian distribution results in significantly lower performance. It is worth noting that the complexity of images, which encodes a high degree and rich information of colors, shapes, textures and spatial relationships between objects, can make it difficult for models to learn effectively from randomly initialized images. On the other hand, using real text sampled from the training set vs. randomly initialized text embeddings does not bring a significant difference. We assume that the pretrained language models are good at generating or transforming 'noise' text embedding into meaningful sentences during the learning process, partly due to the inherent structure and predictability of language. We provide visualizations of real images and 'noise' texts combination below in Fig. 6 and Fig. 7 and Tab. E.1. To our surprise, even though the initialized 'noise' texts are not semantically meaningful to the initialized real images, we discovered a substantial degree of semantic similarity between the initialized real images and the learned distilled text. This suggests the probability of future application of our method in Visual Question Answering (VQA).

Table 9: **Image-Text Pair Initialization.** We compare the retrieval performance achieved with different combinations of image and text initialization strategies. The ✓denotes the use of real images or texts directly sampled from the training set, otherwise indicates the use of randomly initialized image or text, following Gaussian distribution. We can see that if we initialize the image from scratch, the performance will be pretty low. On the contrary, the performance did not drop too much if we start with 'noise' texts and real images, which indicates the importance of image signal for the small distilled set.

		Distillation					
		TR			IR		
Real Image	Real Text	R@1	R@5	R@10	R@1	R@5	R@10
✓	✓	**9.9**	**28.3**	39.1	**4.7**	**15.7**	**24.6**
✓		9	27.2	**40.1**	3.9	13.2	20.6
	✓	0.2	0.7	1.1	0.1	0.5	1
		0.1	0.3	0.4	0.1	0.4	0.8

Figure 6: Initialized Images, iteration = 0.

Figure 7: Distilled Images, iteration = 1000.

'Noise' Texts, iteration = 0.
30 randomly initialized text
from Gaussian distribution, we use nearest neighbor to find their closest sentence in the training set in Flickr30k for visualization purposes.

this man is fit and well toned running enthusiast
the music concert is just started at the giant stadium
a man in a beige shirt and tan slacks sits in a chair next to a hospital patient wearing a blue gown who is sitting cross-legged on his hospital bed
man and woman employed by mongolian barbecue stand at counter
woman cupping water with hands over bathroom sink as child stands beside her
near snowflake sign, man sits while another stands wearing badge and headphones
very brave snow skier doing a flip off a cliff
a guy, dressed nicely, is painting a mural on a wall, with a ladder sitting beside him
dog chasing brown cow and black cow
seems to me looks like people in a work room or office working they all using laptop computers from apple it seems there dr pepper soda and water bottle on
olympian performing on the rings
man walking behind distracted-looking woman carrying bags and camera
three men in caps sit at fireside near cabin, reading at night
the dog with the red collar is white, black, and brown
minor league pitcher
man in chair laughing and talking to others, while handling books
the man, with no shirt, reaches into a bucket to extract the substance inside small brown dog on leash
woman sitting at a park bench reading a book
violin soloists take the stage during the orchestra's opening show at the theater
black dog sitting while eating with neon yellow band around shoulders
several people are standing under a tarp two ladies are facing each other and one has a backpack on with her hands in her jeans pockets while the other one
a boy wearing a flowered shirt raises his arm and jumps
a quarterback is looking to set up a pass from the end zone, while a teammate provides some blocking
a woman in a red coat takes a picture near marble columns at twilight
people with anti-immigration signs
the outside of a restaurant called el triuneo
cheerleaders build a pyramid near the goal-line
baby wears green frog big and makes grotesque face
a yellow, suspended roller coaster on a yellow track is midway through a loop

E.2 Encoder Backbone Selection

In this section, we evaluate the impact of different language/vision backbones on the distillation performance.

E.2.1 Language Backbones

Perhaps not surprisingly, CLIP (Radford et al., 2021) text encoder significantly outperforms BERT in all evaluation metrics, with a striking peak performance in TR R@10 at 92.8% for expert training. This exceptional performance can be mainly attributed to the fact that the pre-trained, off-the-shelf CLIP model is designed to learn a shared embedding space across multi-modalities. Although CLIP also shows a performance drop during distillation, it still retains a relatively high performance recovery ratio. In Sec. G we provide visualization of synthetic data distilled via NFNet and CLIP.

E.2.2 Vision Backbones

The vision encoders carry the main gradient flows for the distillation process. We experimented on several vision backbones, and found that the architecture choice strongly influences the distillation quality. Similar

Distilled Texts, iteration = 1000.
Starting with randomly initialized text from Gaussian distribution, here is the synthetic text after distillation.

```
superhero man leaping in a plaza
a guy in a blue shirt listens to music as he skateboards along the edge of a ramp
tiger woods about to make a putt
little boy pulling a green wagon wearing a sweatshirt and boots
a young girl with blond-hair and glasses sitting at a table in a restaurant
three black young man are working in a semi-deserted area with a pile of construction material and jugs, one of them is
digging
woman buying cups of fruit from street vendor
six men in blue jumpsuits and a man in an orange jumpsuit walk by a shipyard
young girl balances on a reclined man's legs as part of a performance in front of an audience
a woman fillets a fish, as part of preparing a recipe that includes broccoli, celery, and eggs
a young man wearing a white shirt and red shorts kicking a ball
contortionist in strange checkered outfit wearing a white mask
one man plays an acoustic guitar, while another accompanies him on the accordion
male wearing brown shirt holding a microphone with an expression of singing
a young lady in a colorful dress, holds a white stuffed animal stands in the rain hold a plaid umbrella
a person with blue and polka-dot socks jumps on a bed with a red and white blanket
a damaged black color car on the street
skateboarder jumping in air and his skateboard is between his legs
a woman with a guitar sings in front of a building and grass
two woman are sitting on a beach together, facing the water
a busy street with building lined up and people walking down the street outside and nighttime
parasailer doing flip in midair
crowded arena with lots of people wearing yellow, carrying red flags
men in turbans laying down and examining cloth
a woman in a white apron prepares various meats on a large grill
a middle-aged man in a trench coat sleeps on a bus or train
a line of people, some standing and some sitting, are waiting on a platform for a train
three men playing drums, bass, and piano
a dirt-blonde girl in a white top with a key necklace holds a bag, standing in front of a sidewalk of street
cattle-drawn wagons traveling down a paved road and loaded with sticks
```

Table 10: **Ablation Analysis on Language Backbones.** We provide expert training and distillation performance evaluation for both pretrained BERT and CLIP models. CLIP text encoder demonstrates a strong capacity for high-recall retrieval.

Language Model	Expert						Distillation					
	TR			IR			TR			IR		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
BERT	33.9	65.1	75.2	27.3	57.2	69.7	9.9	28.3	39.1	4.7	15.7	24.6
CLIP	**61.2**	**87.5**	**92.8**	**49.8**	**79.8**	**88.3**	**31.4**	**58.8**	**72.0**	**17.1**	**41.9**	**56.2**

to dataset distillation by gradient matching (Zhao & Bilen, 2021b), batch normalization has an impact on the gradient/parameter matching framework. This is mainly because batch normalization incorporates a non-parametric component that can only be accumulated with batches and can not be trained.

Table 11: **Ablation Analysis on Vision Backbones.** We provide an extensive evaluation of several pretrained vision backbones including NFNet_l0 (Brock et al., 2021b), NF_ResNet50 (Brock et al., 2021a), NF_RegNet (Xu et al., 2022), and ResNet50 (He et al., 2016). This underscores the influence of architecture and highlights the potential negative impact of batch normalization for distillation.

Vision Model	Expert						Distillation					
	TR			IR			TR			IR		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ViT (LoRA)	**40.7**	**69.8**	**80.1**	**28.8**	**59.3**	**73.4**	**10.4**	23.6	38.7	**5.4**	**18.8**	**27.4**
NFNet-l0	33.9	65.1	75.2	27.3	57.2	69.7	9.9	**28.3**	**39.1**	4.7	15.7	24.6
NF_ResNet50	28.9	56.6	71	22.8	50.1	63.4	6.5	18.2	28.1	3.5	11.6	18.7
NF_RegNet	26.9	57.2	70.2	21.1	50.1	62.9	7.8	21.9	33.3	3.3	12.7	20.5
ResNet50	18	43.5	59.5	13.4	36.6	49.9	0.5	2.4	3.8	0.3	1.6	3.6

E.3 Pretrained vs. Non-pretrained

Tab. 12 demonstrates the pretraining influence of the backbone encoders. Optimal performance is observed when both language and vision backbones are pretrained. This emphasizes the importance of pretraining before the expert training stage for large models and datasets.

Table 12: **Pretraining Impact.** Expert performance comparison for different pretraining configurations of vision and language backbones. The checkmark (✓) indicates the model was pretrained.

		Expert					
		TR			IR		
Language Backbone	Vision Backbone	R@1	R@5	R@10	R@1	R@5	R@10
✓	✓	**33.9**	**65.1**	**75.2**	**27.3**	**57.2**	**69.7**
✓		4.4	14.1	20.7	3.5	11.4	18.8
	✓	0.5	1.1	1.8	0.3	0.7	1.4
		0.3	1	1.5	0.1	0.7	1.3

E.4 Synthetic Steps

The synthetic step size plays an important role in optimizing the dataset distillation performance, as shown in Tab. 13. Using larger synthetic steps tends to achieve better distillation performance.

Table 13: **Synthetic Steps Impact**. Larger synthetic steps greatly improve performance. For 100 pairs with a synthetic step of 1, the performance is even below random selection. Setting the synthetic steps to a low value typically takes longer to optimize the distilled set and it is challenging with very small sets (e.g., # Pairs=100).

		Distillation					
		TR			IR		
#Pairs	#Syn Steps	R@1	R@5	R@10	R@1	R@5	R@10
100	1	0.5	2.1	4.4	0.3	1.5	2.8
	2	7.1	23.4	32.9	3.0	10.2	16.4
	4	8.2	24.9	35.2	3.5	12.2	20.7
	8	**9.9**	**28.3**	**39.1**	**4.7**	**15.7**	**24.6**
200	1	3.2	9.3	14.1	1.6	5.2	8.8
	2	6.5	19.2	29.1	1.6	5.9	10.0
	4	8.2	24.5	34.4	2.2	7.4	11.8
	8	**10.2**	**28.7**	**41.9**	**4.6**	**16.0**	**25.5**
500	1	6.6	18.1	25.5	2.1	10.1	16.3
	2	8	21.7	31.3	3.8	14.9	23.2
	4	8.1	23.6	34.9	4.4	15.2	23.7
	8	**13.3**	**32.8**	**46.8**	**6.6**	**20.2**	**30.0**
1000	1	7.3	20.6	29.7	3.9	13.2	20.7
	2	8.8	26.8	36.6	5.7	17.4	26.4
	4	10.4	29.1	37.9	6.6	19.5	29.5
	8	**13.3**	**34.8**	**45.7**	**9.1**	**24.1**	**33.8**

F Beyond Trajectory Matching

In this section, we further provide experiment results of a distribution matching (Zhao & Bilen, 2023) baseline adapted to the vision-language setting. To use distribution matching for vision-language dataset distillation, concretely, we minimize the maximum mean discrepancy (mmd) between two distributions by sampling NFNet with different initialization and pretrained BERT. Similar to the distribution matching setting for image classification, we update the distilled data via mmd for vision and language modalities to match the original data distribution in a family of embedding spaces. We provide the comparison of Our method w/ DM (distribution matching) and Our method w/ TM (trajectory matching) on Flickr30K (R@1) in Tab. 14.

	TR		IR	
# pairs	**Ours w/ DM**	**Ours w/ TM**	**Ours w/ DM**	**Ours w/ TM**
100	3.2 ± 1.8	**9.9 ± 0.3**	1.4 ± 0.7	**4.7 ± 0.2**
200	3.3 ± 1.3	**10.2 ± 0.8**	1.4 ± 0.4	**4.6 ± 0.9**
500	5.8 ± 1.5	**13.3 ± 0.6**	4.1 ± 0.9	**6.6 ± 0.3**
1000	6.1 ± 2.7	**13.3 ± 1.0**	4.9 ± 1.8	**7.9 ± 0.8**

Table 14: Comparison of our method using Distribution Matching (**Ours w/ DM**) and Trajectory Matching (**Ours w/ TM**) on the Flickr30K dataset. The table shows retrieval performance (R@1) across different numbers of pairs. The results indicate that our method with trajectory matching consistently outperforms distribution matching, particularly in scenarios with smaller data budgets.

Looking forward, we hope our method could serve as a roadmap for future studies exploring more complex settings with new state-of-the-art (SOTA) methods. New SOTA dataset distillation methods can adopt low-rank adaptation matching to scale efficiently with large and complex models, and can incorporate bi-trajectory co-distillation to handle textual data more effectively. By doing so, these methods can extend their applicability to previously infeasible models for distillation, such as those involving ViTs, thus improving the scalability and efficiency of the distillation process. New approaches that distill from both text and image data can consider using methods similar to bi-trajectory matching with contrastive loss to learn the interactions and redundancies across multimodalities.

G Additional Visualizations

Here we include a number of visualizations of the data we distilled from the multimodal dataset (both Flickr30K Tab. G and Fig. 8, 9 and COCO Tab. G and Fig. 10, 11) for a more intuitive understanding of the distilled set. We provide 50 distilled image-text paired examples including their visualization before the distillation process. Unless otherwise stated, these experiments are conducted using 100 distilled pairs, with pretrained NFNet (Brock et al., 2021b) and BERT (Devlin et al., 2018) as backbones and the synthetic step is set to 8 during distillation. We provide visualization of distilled data using NFNet and CLIP in Tab. G and Fig. 12, 13 in the end.

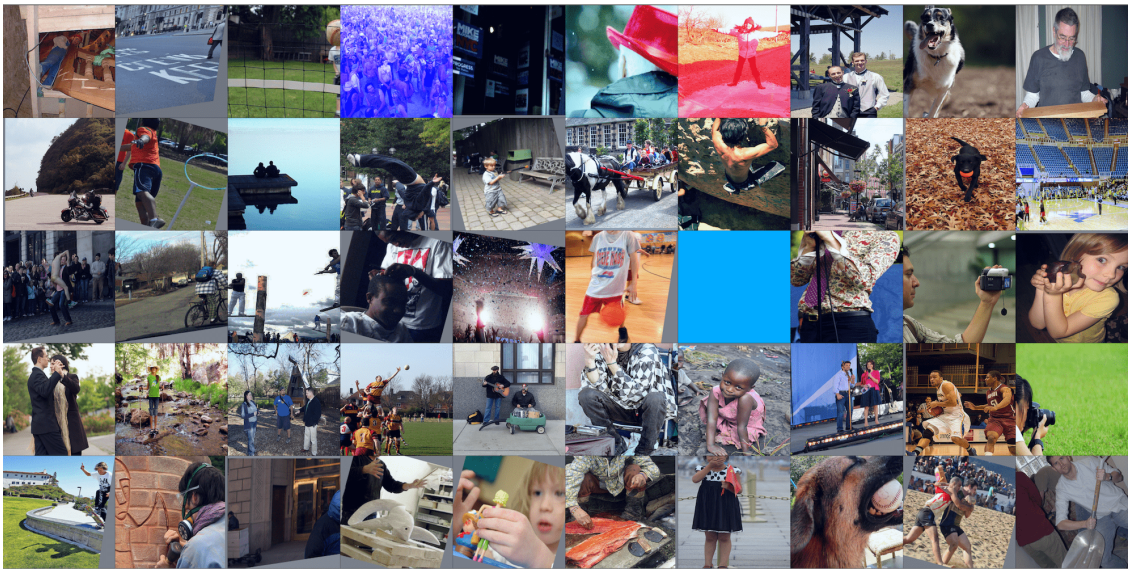

Figure 8: Flickr30K Initialized Images, iteration = 0.

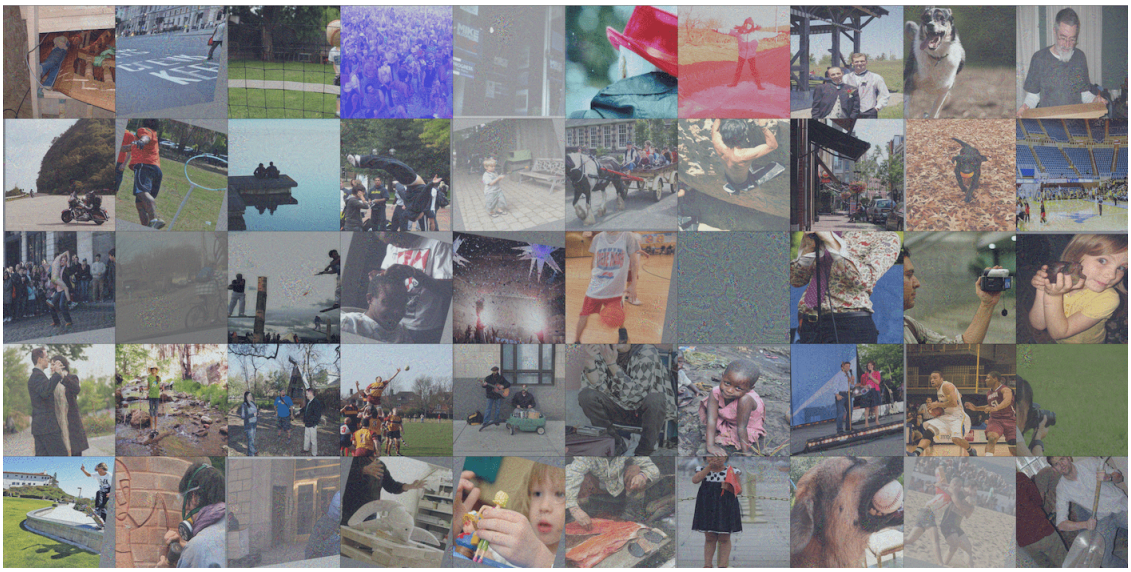

Figure 9: Flickr30K Distilled Images, iteration = 2000.

Flickr30k Initialized Texts, iteration = 0.

a construction worker stares down a flight of stairs being built
a man in a suit walking across a city street
a child hits a baseball into a net
a large crowd of people
an old man, behind him on glass is many political advertisements
an old man with white hair in a red hat
a young girl, on a road, playing on the ice
two men dressed up before a big event
a dog is running through a field with its tongue hanging out
an older man in a gray shirt with a white long-sleeve shirt under it holding up a small wooden cabinet
a motorcycle is parked along side a mountain road while another goes down the road
a man in an orange shirt and black shorts throws a ball through a hoop while another man watches
two people sit on the end of a dock
a man performs a back flip while preparing for an outdoor performance or competition
a little boy plays with a toy gun
a bunch of young children are riding on the back of a trolley while being carried by a black and white horse
a man climbing up on a rock ledge
a young man sits on a bench in a downtown setting
a black dog carries an orange ball, walking on the ground covered in leaves
basketball players practicing for their game
a man in just his underwear jumping on a man surrounded by a crowd of people
a man wearing lots of plaid riding a bike through the streets
men are trying to cut down trees
a black man getting a haircut
this was a big new years event the people were sing and dancing all night
a boy wearing a steve nash shirt dribbles a basketball on an indoor court
a series of men taking a break from riding their motorcycles
a brown-haired man in a patterned shirt and purple tie is singing into a microphone
a curly dark-haired man holds a small camcorder and films in a person in front of him
a young girl in a yellow shirt holds a rather large snail in her hands next to her cheek
two young men dancing in the street
a girl standing in a shallow creek, wearing stilts
3 people standing on a park talking to each other
a group of young men in colorful uniforms playing with a white ball
two guys are on the side of the street playing a guitar and drums
a mime applying his makeup
a child decorates a shoe with colorful sticks
a man and a woman are up on a stage with microphones in their hands
two basketball players on opposing teams, one in white, the other in red, are mid-game, running down the court, white-uniform player with ball-in-hand
one young lady with black hair in a ponytail wearing a black bracelet and a white shirt, taking pictures with a black camera that has a shoulder strap laying in
a boy on a skateboard is on a wall near the water and next to grass
a sculptor is carving a picture of a knight into a brick wall
a man in a blue coat is walking on the sidewalk
an old man wearing glasses is holding a stick
child playing with doll-like toy
an old man wearing a hooded sweatshirt is crouched over a fish that has been cut open
a young girl in a black dress is holding a red flag and covering a happy expression
a brown dog with a baseball in its mouth
man in white and red tackling man in green shirt for the ball
a man in a white t-shirt is holding a snow shovel

Flickr30k Distilled Texts, iteration = 2000.
construction workers repair walls of a subway
a ship in a harbor at night with a city skyline behind
baseball pitcher throwing a pitch
group of people sitting around a table for a meeting
a man points to something as he is talking to a woman wearing white pants, as they stand in front of a store
man in red sweater with a backwards hat
women wearing winter coats crossing the street next to parked cars and walking down street
the bridal party poses with the bride and groom, all wearing black except for the bride
an old lady, wearing a red hat, is standing on the sidewalk of a park
a man grilling hotdogs and sausages
a motocross bike kicks up dirt as it is being ridden around a bend in the circuit
nine women in blue and purple dresses and one man wearing a purple shirt and black pants, clap while a man dressed in black dances
two young men and two boys are sitting down on a boat next to an anchor and watching the water
while playing soccer, a man in yellow starts to fall, while a man in white trips over him, stepping on his ankle in the process
a little boy is walking on a fallen tree in the woods
a jockey and horse in the middle of other jockeys and horses during a race, in the middle of jumping over a hurdle
an extreme man snowboarding up side down a mountain
one man, in a blue jacket, is sitting in the rain under a green umbrella
the brown and white dog is running to catch something
boy takes a bath with diving mask and snorkel
angry looking businessman walking down sidewalk
a person on a bmx bike, leaping onto a bench
two people and a dog are in the snow
young shirtless boy sleeping on a couch with his hand on his chest
a person spins a sparkler around at night and sparks fly through the air and across the ground
a woman, wearing sunglasses, a red athletic top, and running shorts competes in a marathon
a ballet dancer wearing a blue tutu doing the splits, mid-leap
woman on street corner smiles and talks on her cellphone
policeman taping off an area by a group of firemen
a man with a beer and another man facing each other, talking
a woman and two children reading outside on a stone bench
man on motorcycle riding in dry field wearing a helmet and backpack
a man in a white shirt and black exercise shorts walks on a sidewalk, which is located behind a street under construction and in front of a two garage house
a man is hitting a golf ball out of a sand trap, there are green grass all around him
some young adults are playing saxophones and clarinets outdoors
a young man sitting on a rock on the shore of a body of water looking contemplative
a young boy, covered in mud, plays on the beach
shirley manson poses in front of a microphone on stage while holding a large blue, red, and white flag behind her
two hockey players playing offense and defense
a group of friends, 3 boys and 2 girls, jump in the air holding hands for a photo
a boy skateboards and does a jump over another skateboard
ginger baby playing with a train setup made out of counterfeit lego
a group of choreographed rollerskaters dancing
little blond girl in her jacket sticking out her tongue while holding a red balloon
a blond little girl wrapped up in a pink care bears blanket
an oriental man, wearing a white shirt and apron is cooking
one boys sits on a giant mortar gun as another boy walks toward him
brown dog trying to bite a white ball with yellow, green and blue puppy toes
woman standing on the shore of a beach
2 males, one in a red shirt and one in a yellow shirt, cleaning a room

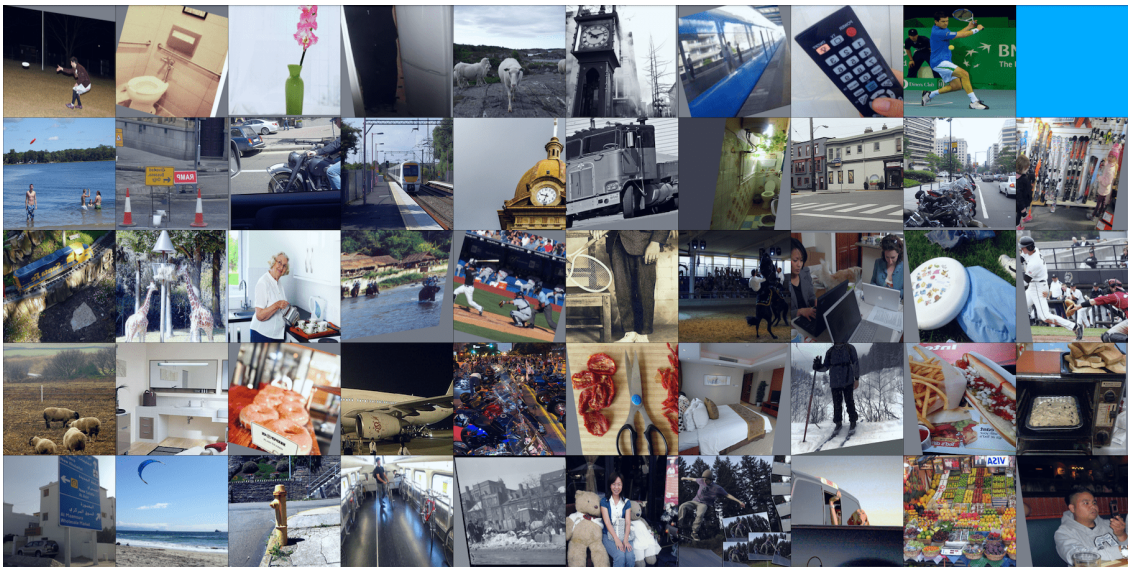

Figure 10: COCO Initialized Images, iteration = 0.

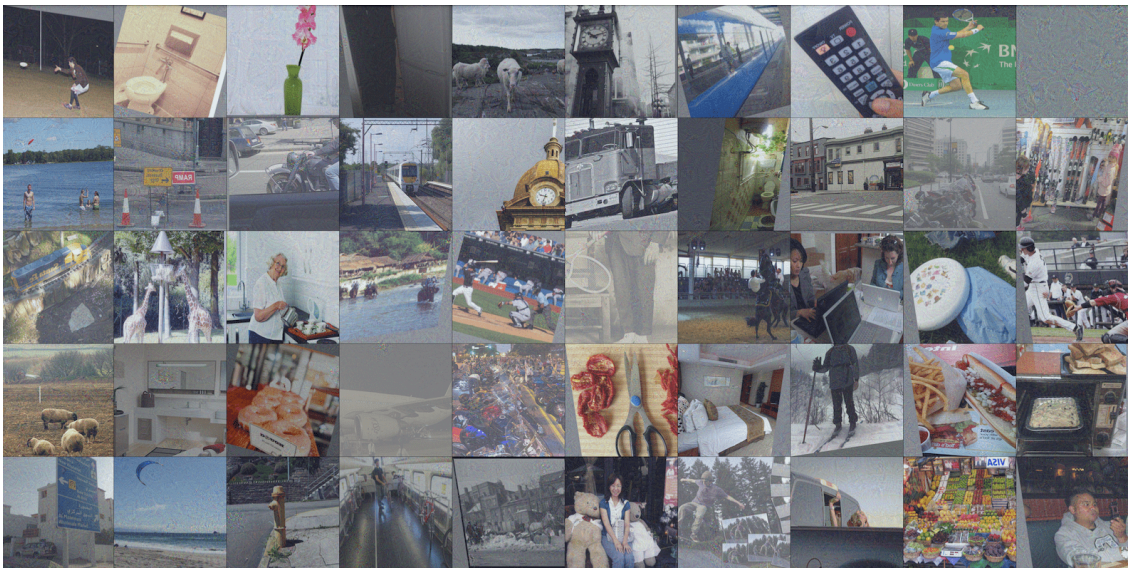

Figure 11: COCO Distilled Images, iteration = 2000.

COCO Initialized Texts, iteration = 0.
a photo taken at night of a young man playing frisbee
a bathroom toilet is surrounded with silver handrails
a pink flower is sticking out of a green vase
a woman poses next to a fridge
a small group of sheep on the coast
a black and white photo showing a large clock tower on a building
the people are waiting to be picked up
a hand holding a black television remote control
a man swinging a tennis racket at a tennis ball
a man holds a small animal to his face
three people are in the water as a frisbee is in the air
yellow and red street signs warning larger vehicles in a large city
a man riding on the back of a motorcycle
the train is traveling down the railroad tracks
a building with a large clock on it
a man standing on a wheel of a big tarck
a small bathroom with a toilet, sink and window
a florist shop and other buildings on and near a street corner
a line of motorcycles parked on a street
two young girls in a store looking at skis
a small toy model train on a track
a couple of giraffes are outside in the wild
a woman pouring coffee into cups on a counter
people riding on top of elephants across a river
a hitter swings at the baseball and misses
an old picture of a guy holding a tennis racquet
there are people watching three men on horseback
two people sitting at a table with laptops
there are many things laying on the ground
a batter swings hard at a low ball as the catcher reaches out his glove
five sheep stand around a large dirt field
a bathroom with a sink, mirrors and chair
glazed donut sitting on a wooden table top in a donut shop
a huge chinese aircraft is sitting at an airport with people unloading
a bunch of motorcycles are parked together outside
cutting board with scissors and dried food on it
a clean, decorated bedroom is pictured in this image
a man standing on skis at the top of a hill under high tension wires
a long hot dog and french fries are on a plate
a pan of dough is going into the dirty toaster oven
a road sign with both english and arabic directions
a kite being flown on the beach while people watch
a yellow fire hydrant is on the corner of an old sidewalk
a man with a bicycle and a helmet on his head in a subway car
a horse struggles to draw a loaded cart through piles of snow
a woman is sitting between two large teddy bears
a kid skateboarding while other kids stand and watch
there is a dog in the back of the truck
a large assortment of fruits lie on display in a market
he’s taking a picture of his friends at the restaurant

COCO Distilled Texts, iteration = 2000.
dog flying in mid-air running after a frisbee
bath tub with metal shower head, vanity mirror, and small utilities compartment
a white vase with pink flowers and large green stems
this apartment has an kitchen with a refrigerator, stove, dishwasher, and cabinets
a ski resort with many people gathered around the outside lodge
grandfather clock hanging on wall next to a grandfather clock
the kitchen is brightly lit from the window
someone is playing with a nintendo wii controller
a woman swinging a tennis racket, while standing on a tennis court
a jet plane taking off into the air
a sign warning people to stop at red lights
man swinging baseball bat with ball in air and crowd watching
a man hitching a trailer with water sports equipment to a sports utility vehicle
four trains lined up at the train station
a large tower has a clock towards the top
people standing at a bus stop about to board a bus
a small bathroom with a sink a mirror
man admiring a motorcycle in parking lot, near a large building
a motorcyclist in a red and white suit riding a red and white motorcycle
a little boy playing tennis on a tennis court
a locomotive train resting on some tracks next to a building
two giraffes graze on some tall plant feeder
woman looking at camera while lying in bed
a herd of adult elephants with a baby elephant waling through a forest
baseball batter in a wide stance, waiting for a pitched ball
the cars were parked along the street near the traffic light
the travelers are getting around by horses
a cluttered desk with a laptop opened to flickr
the lady is sitting on the wood bench
a baseball player is preparing to swing a baseball bat
two lambs eating hay from ground of a field
some brown cabinets a black oven a tea kettle and a microwave
the reception ifs full of professional people
baseball batter ready to strike arriving ball and umpire waiting to catch if he misses
there lot of motorcycles park in a line with a white car, a red car and a van park not far from the motorcycles while there is man riding on
a very clean room and a pair of scissors
the bedroom has a wood closet and bookcase near the bed
a line of skiers heading towards a cabin in the mountains
a plate topped with onions rings next to a hamburger and hot dog
the personal sized pizza has been topped with vegetables
a sign letting people know about the castle rising castle
girl and two males standing on a beach
there is a white fire dyrant on the corner street
a man skateboarding on street in front of a bus
vintage black and white photograph of two baseball players
raggedy ann doll sitting in a chair with a pooh bear
blonde haired boy doing a jump while riding a skate board
a dog driving a car down a street
there are bananas, apples and oranges in the bowl,
a stop light with the green light lit

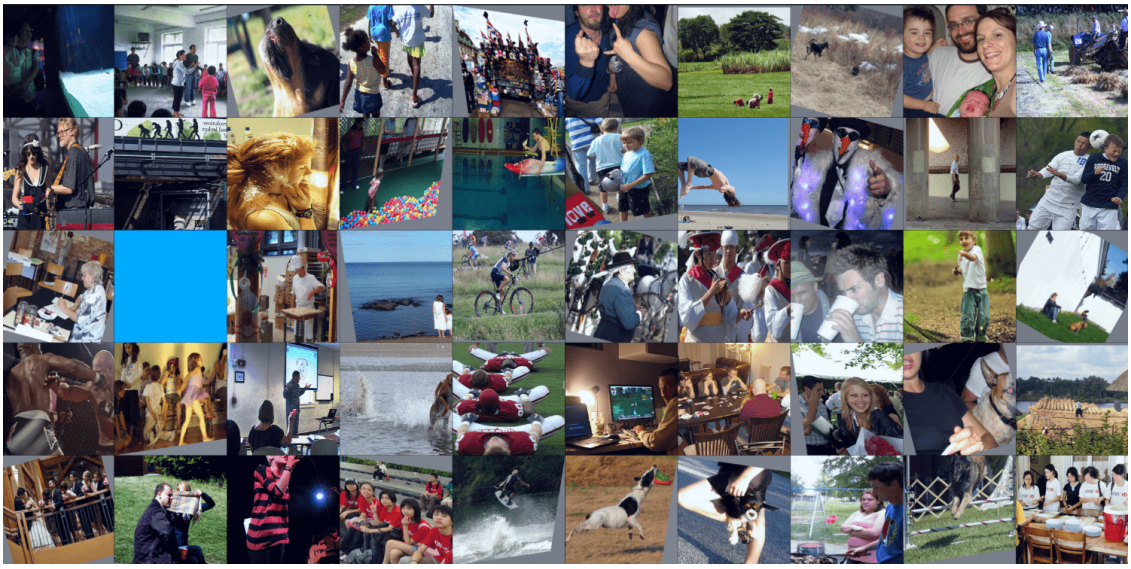

Figure 12: **CLIP**, Flickr30K Initialized Images, iteration = 0.

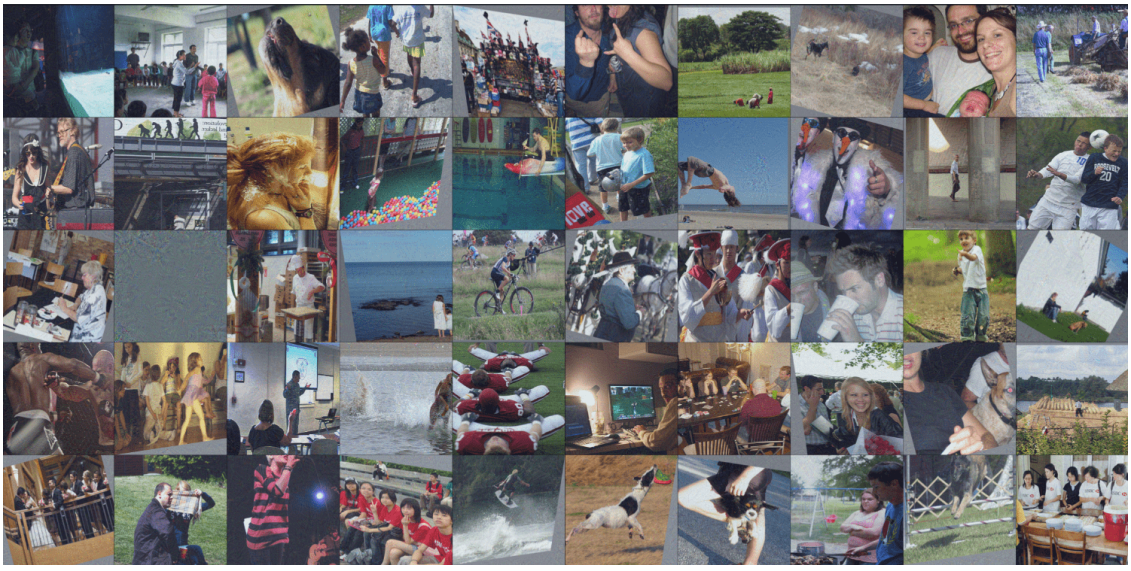

Figure 13: **CLIP**, Flickr30K Distilled Images, iteration = 2000.

CLIP, Flickr30k Initialized Texts, iteration = 0.
a woman holding a child is standing in front of a tank
a woman and man with a child in the center of a circle of children
a black and brown dog eyeing a fly
four kids next to a blue house are walking down a street
tourists look at merchandise on a street vendors display in england riffling through cards and maps
two people wearing blue clothing are making hand gestures next to one another
four workers in a field harvesting
a dog approachs a small creature in a barren, snowy field
a woman holding a baby and a man in glasses holding a small boy smile at the camera for their family photo
four men are harvesting a crop on a farm
two musicians on stage in front of a microphone stand
a man in a suit is walking under a metal bridge
girl getting her hair dyed
a child is in a ball pit while three adults watch her
a man is sitting in a small kayak on a diving board
a boy in a blue shirt holds a toy helmet in his hands while standing on a path in a park
a boy doing a flip in the air at the beach
two men are dressed up as snowmen
a man walking underneath a bridge glances at the camera
two soccer players are about to butt heads getting the ball
a woman is at an art studio, painting a mural from her art supplies
a boy wearing a black wetsuit stands on a crowded beach
a chef prepares a table with food behind it
two blond girls in white dresses, one much smaller than the other, stand on the bank of a large body of water
a man on a black mountain bike rides through a course filled with other bikers
a man in a hat stands with decorated horses around him
an asian marching band with uniformed members in beige, yellow, and red play in the street
two men with angry faces drink out of white cups
young boy plays with leaves in a green wooded area
a person siting against a wall with a dog
one fighter delivers a hard blow to the face of another fighter
a young group of children sitting in a row against the wall
a teacher stands in front of a projector and a student at the front of the class has her hand raised
a dog is running in a large body of water causing it to splash
football players are stretching together
man playing video game instead of working
a group of people at dining table playing a card game
young woman celebrating her graduation
a dog looks on as a woman eats
a man wearing a flannel shirt and black pants is working on a new reed roof on top of a house
newly married couple having their first kiss
a woman plays hide-and-go-seek with a check scarf as she sits with a man in a dark colored jacket
a woman with short blond-hair smiling and singing into a microphone while a man in a striped shirt, further back, plays an acoustic guitar
a group of asian teenagers wearing red t-shirts sit on steps
a man gets lots of air time as he wakeboards
a black and white dog is running through the field to catch something in its mouth
a person with a small dog caught in her legs
a young man tends chicken wings on a barbecue while a young woman in a pink shirt watches
black and brown dog jumping over hurdle with white supports
a group of women wearing shirts that say, hsbc is standing by a table with food on it

CLIP, Flickr30k Distilled Texts, iteration = 2000.

a young asian girl petting an animal of some sort and the animal is laying down enjoying it
a group of men in ethnic dress are dancing
brown and tan dog, mouth open with tongue hanging out, running in the grass
an older black woman wearing a colorful print dress stares directly at the camera
woman in red shirt shopping in a outdoor market
a woman in a blue shirt with no bra
a man holding a bag walking down a long staircase
a man in black walking down a street
a woman holds a baby in a blue jumper
a small car in an open field
a group of male musicians are playing instruments including guitar and drums
a man is standing inside a subway train with his mouth wide open
a barber shaving someone's head
a group of people are shopping in what looks to be a christmas store filled with colorful toys
a man in high rubber boots and a plaid shirt is pushing a broom over the mossy blacktop
three males with cameras near each other, two sitting and the third standing, in what might be a park during a sunny day
a girl jumping up in the air with her hands above her head
two people, one of whom is in a santa costume, pose wearing funny glasses
a group of people walking through an alley along a cobblestone street, between two buildings
a soccer player in a green jersey kicks a blue and yellow ball
a man painting over graffiti
two dogs running near a river while one dogs in swimming in it
a man in a black hat looks surprised at the deli counter
a woman sits in a chair on the beach, backdropped by the ocean
a young man popping a wheelie on his bicycle while riding down a country road
a person attempts to rope a black cow while riding a horse
men dressed in red and white playing musical instruments
two men drink beer out of tall drinking glasses
a little boy is eating on a sidewalk
a man is standing inside a doorway that is in a wall painted with a mural of a woman
a man wearing a hat has his eyes closed, as another man in a red shirt is licking his face
a family of 3 sits and poses on a couch together
a young boy in a sports uniform stands in front of a group of children
a little boy plays outdoors in water spurting up from an inground fountain
two men in red pants do acrobatics with a ladder
a young caucasian man sits at a desk using a laptop computer
a group of people is sharing a meal at a large table at a restaurant
a group of people protest with one holding up a cardboard sign
a group of people and their dogs at a dog show
a man with a plaid shirt is working on some wood in his workshop
the fellow in the black suit at a formal occasion has a salmon rose in his lapel
a man and a woman walking across a field of grass
a woman singer holding a microphone singing at a concert
young asian female sitting in a pose on a stone wallas she is being photographed
the man is standing by a creek in blue flannel shorts
a black and white dog leaps to catch a frisbee in a field
a man is feeding two exotic birds
an asian chef is in the foreground, working over a steaming grill while a younger man is behind him
a young man is skateboarding down the railing of some stairs
a female chef examines a piece of bread while showing it to the camera