Filtering, Distillation, and Hard Negatives for Vision-Language Pre-Training

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende
Yash Patel, Yi Wen, Vignesh Ramanathan and Dhruv Mahajan

Poster ID: TUE-PM-271

Meta

Team


Filip Radenovic


Abhimanyu Dubey


Abhishek Kadian


Todor Mihaylov


Simon Vandenhende

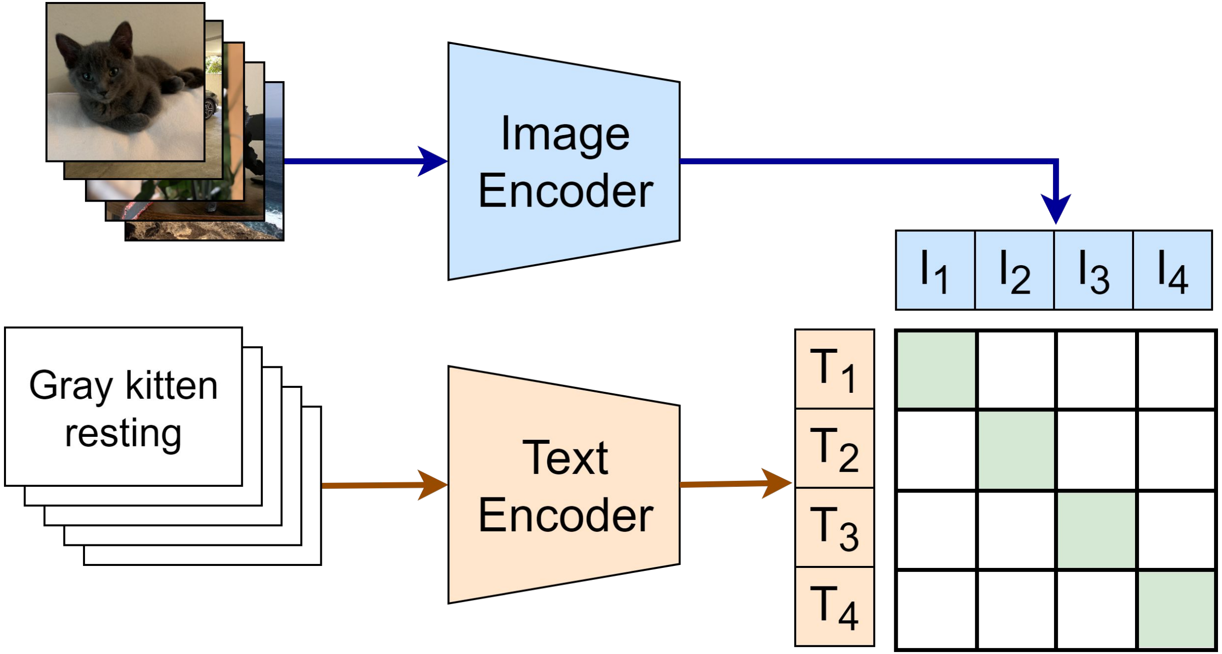

Yash Patel


Yi Wen


Vignesh Ramanathan


Dhruv Mahajan

Dual encoder vision-language pre-training



[CLIP, Radford etal, 2021]
[ALIGN, Jia etal, 2021]

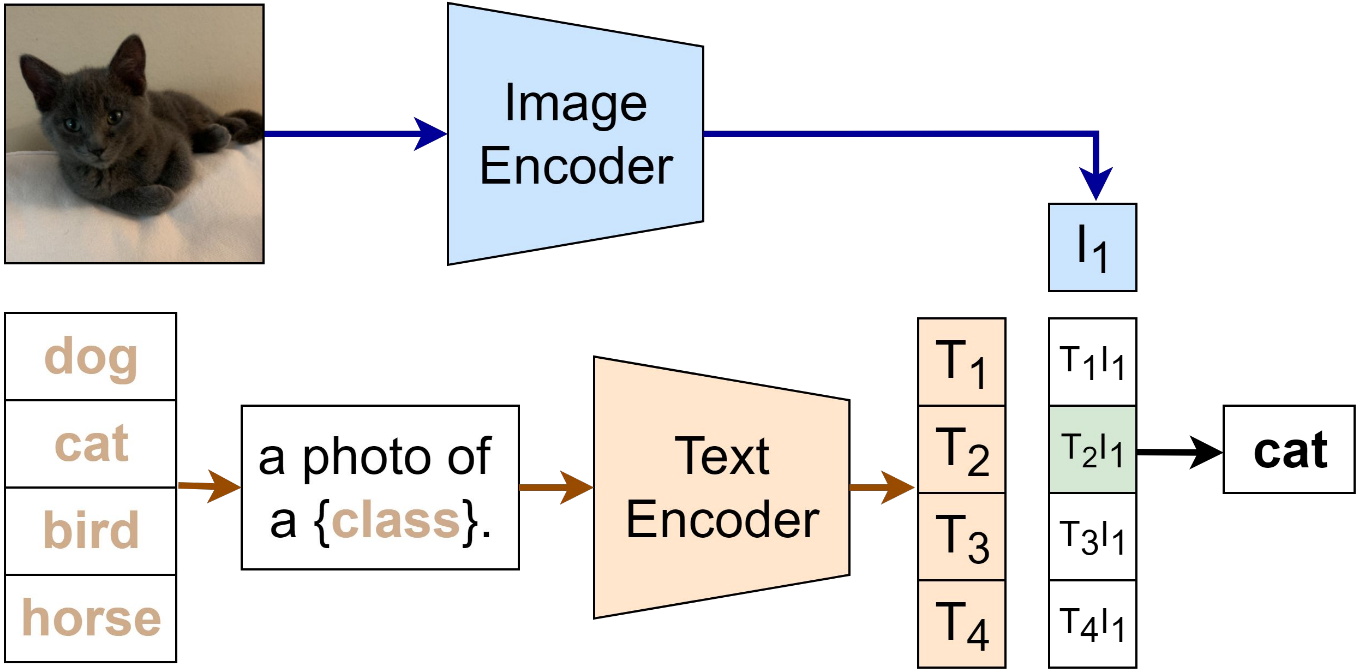
Training data
(image-text pairs)

- CLIP: 400M (private)
- ALIGN: 1.8B (private)

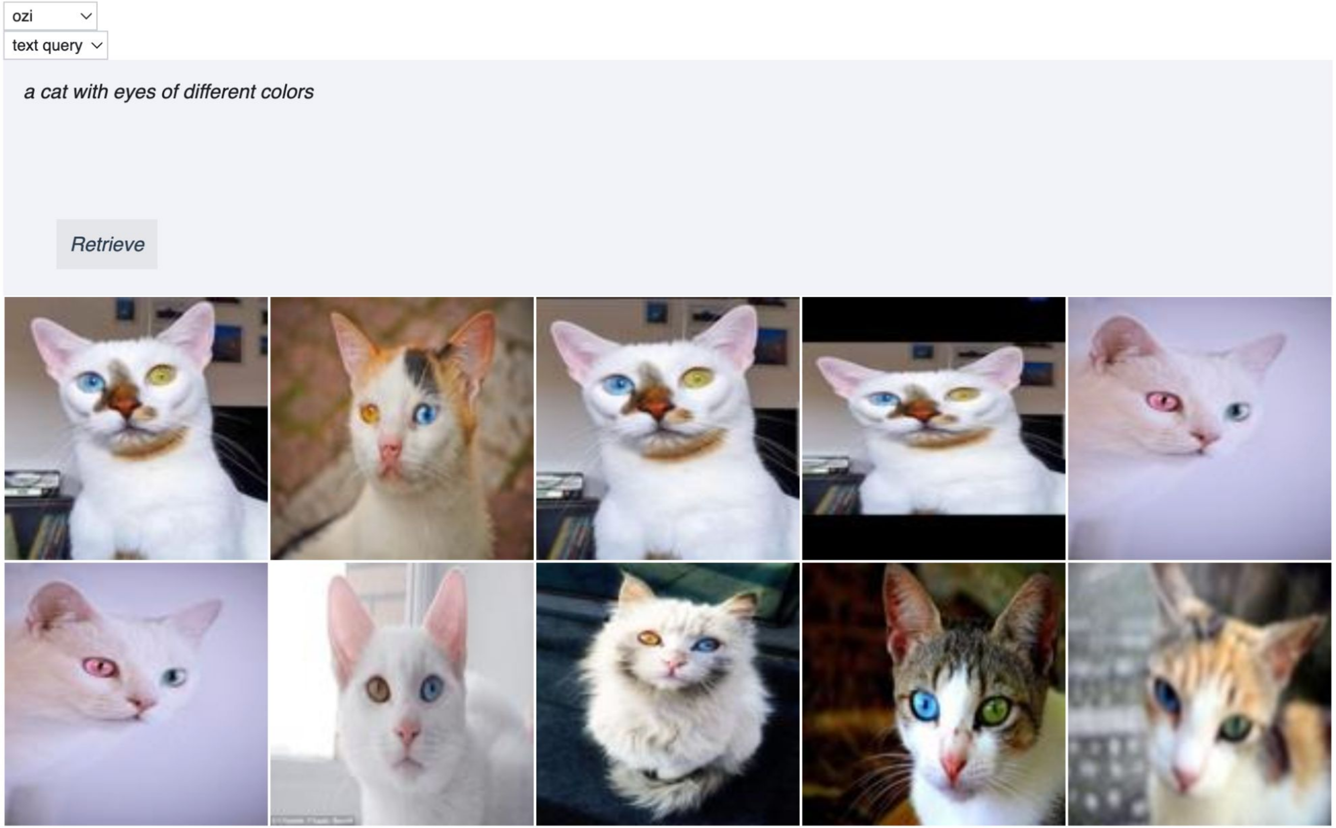
- **LAION: 2B (public)**

[github.com/mlfoundations/open_clip]

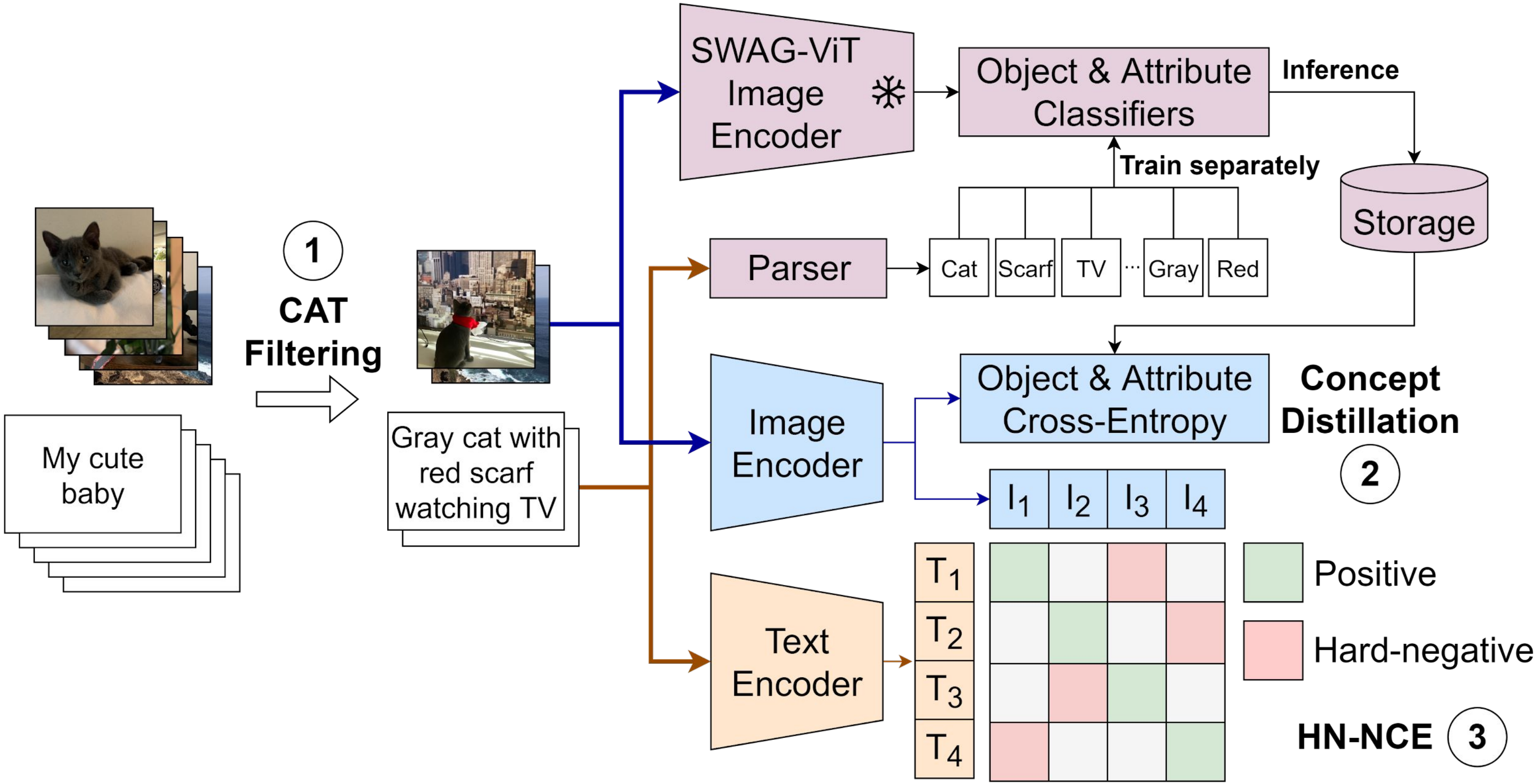
Zero-shot image classification



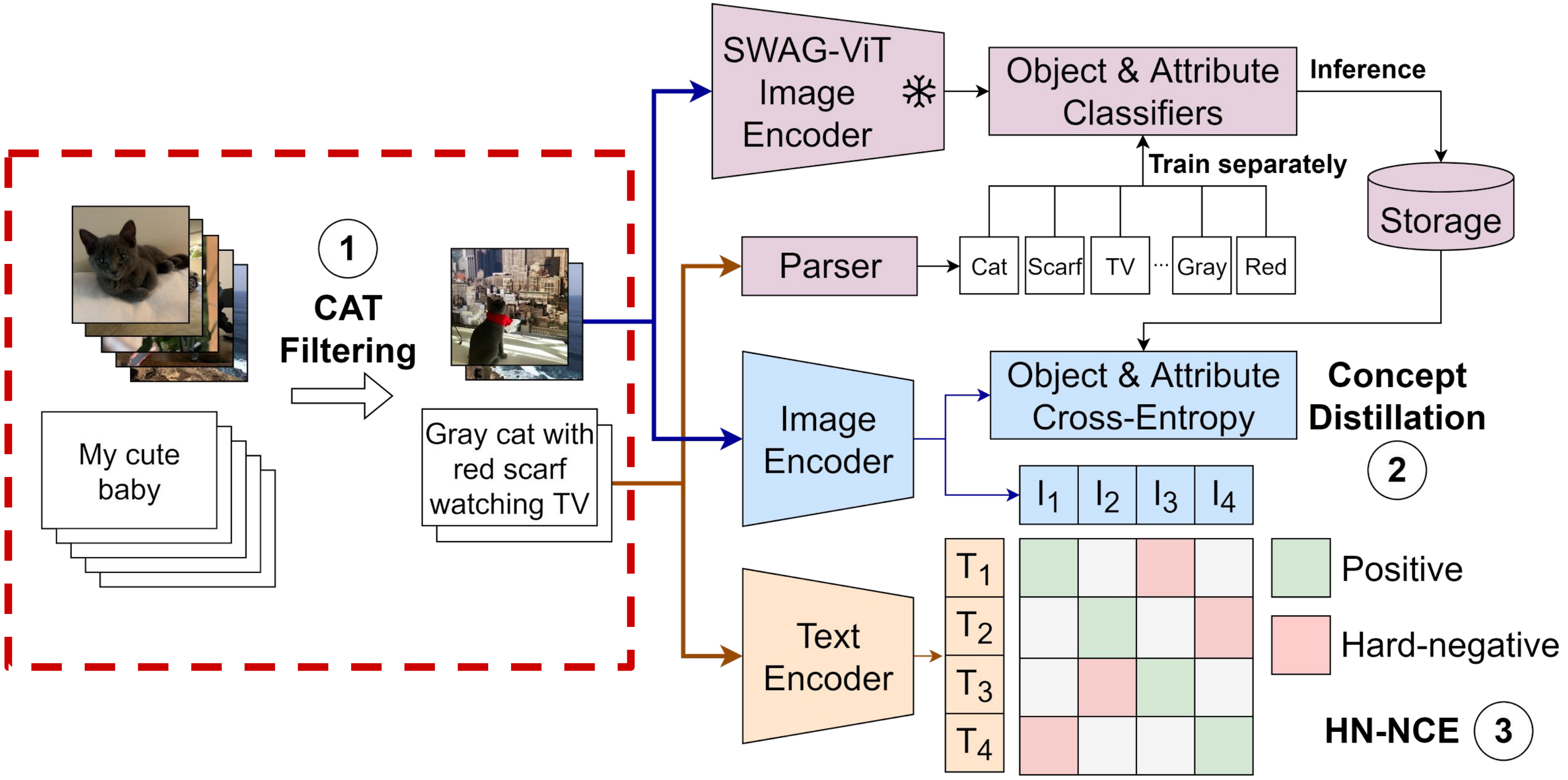
Text-to-image retrieval



Our dual encoder framework



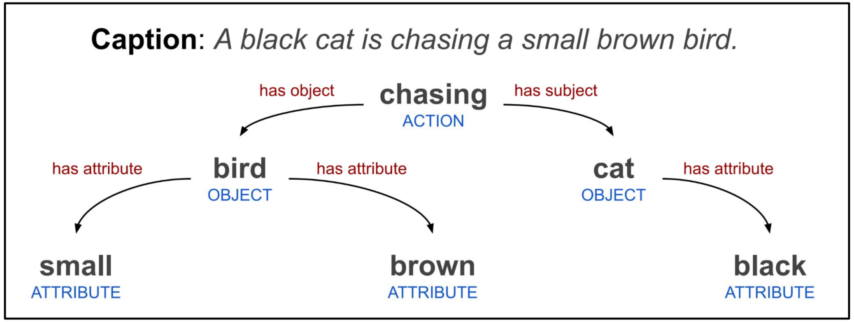
(1) Complexity, Action, and Text (CAT) filtering



(1) Complexity, Action, and Text (CAT) filtering

We filter noisy LAION-2B web dataset based on:

- **Complexity:** keep if at least one relation to any object present in the parse graph



- **Action:** keep if at least one action present in the parse graph

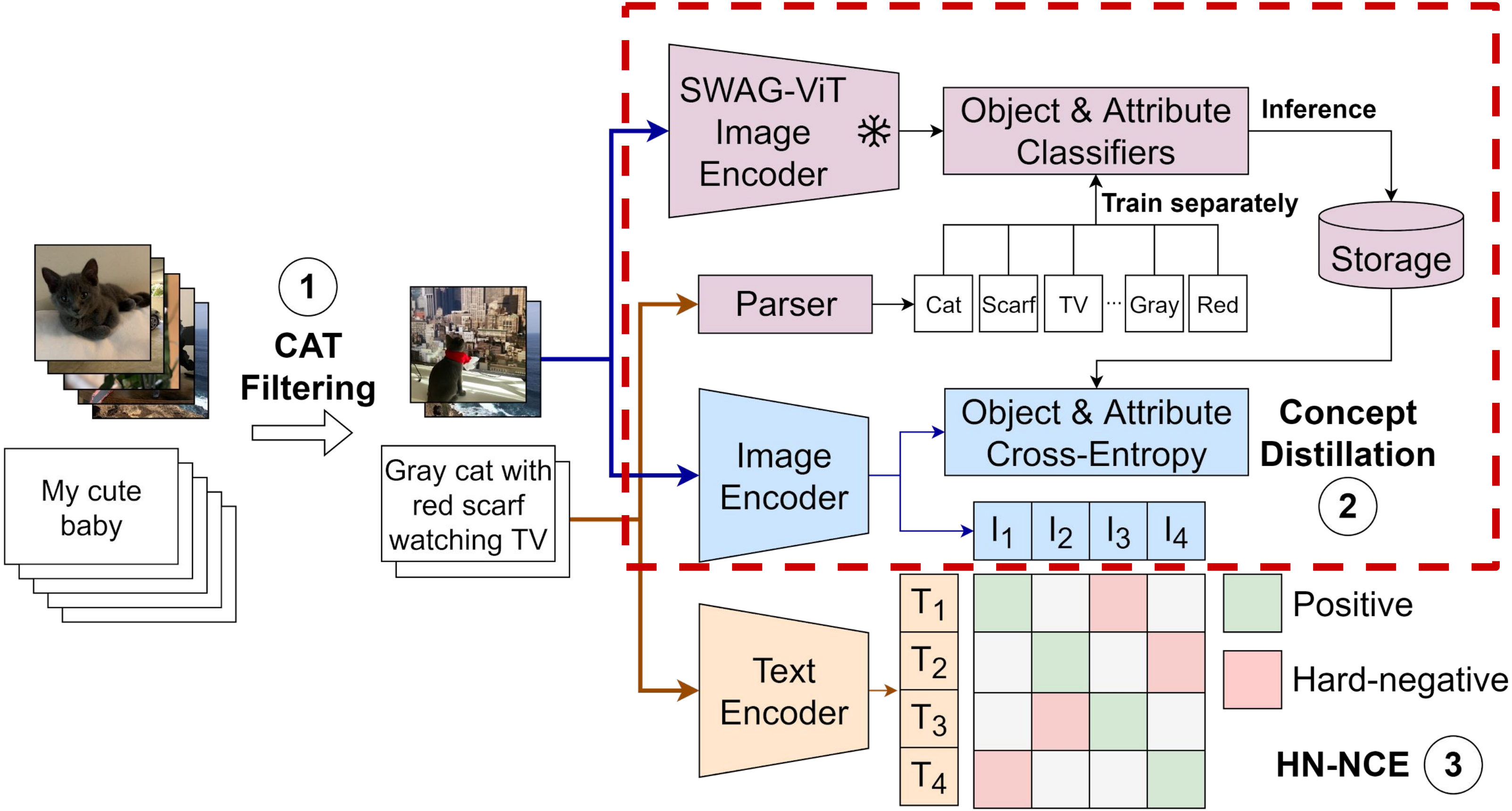
- **Text:** remove if caption present in the actual image

(1) Complexity, Action, and Text (CAT) filtering

Table 1. Evaluating effect of using LAION-2B subset filtered on complexity (C), actions (A), and text-spotting (T). CLIP denotes filtering pairs with CLIP score bellow 0.35. Evaluation performed on ViT-B/32 model architecture trained for 4B processed samples.

#	Filter				Size	IN	COCO		Flickr	
	CLIP	C	A	T			T2I	I2T	T2I	I2T
1					1.98B	60.8	33.7	52.1	59.3	77.7
2	✓				440M	52.5	29.8	46.1	54.8	72.0
3		✓			1.71B	60.8	33.9	52.5	60.8	77.8
4		✓	✓		642M	58.7	35.9	53.8	64.3	82.0
5		✓	✓	✓	438M	**61.5**	**37.6**	**55.9**	**66.5**	**83.2**

(2) Concept Distillation



(2) Concept Distillation

1. Parse image captions using a semantic parser that extracts objects and attributes from text and use these as pseudo-labels.
2. Train the linear classifiers on the teacher model embeddings with a soft-target cross-entropy loss, after square-root upsampling low frequency concepts.
3. Use these trained linear classifiers to generate two softmax probability vectors - for objects and for attributes, respectively.
4. During multimodal training, we use the cross-entropy loss with these pseudo-label vectors as targets.

(2) Concept Distillation

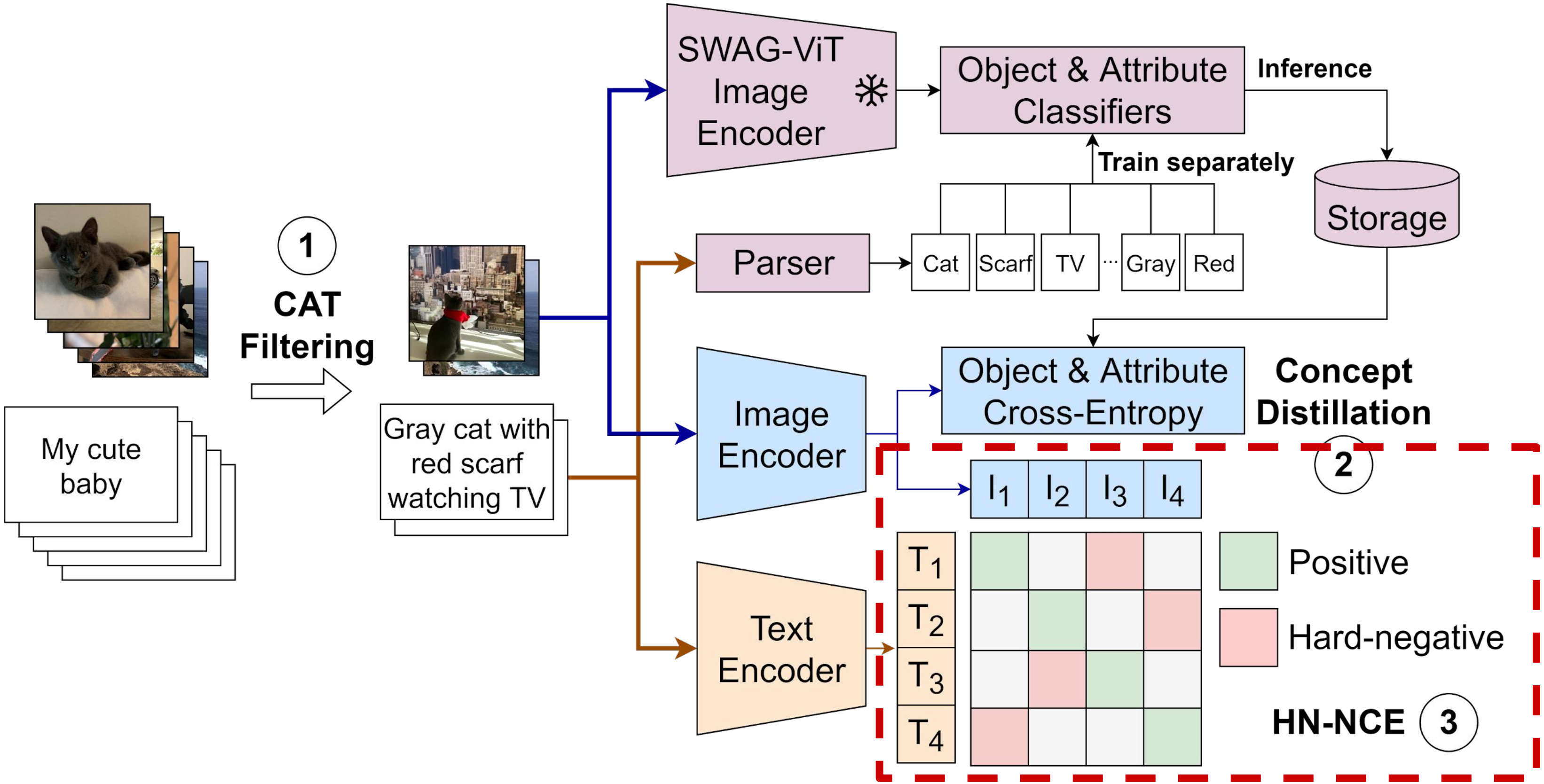
Table 2. Evaluating effect of using different initialization or distillation approaches. Evaluation performed on ViT-B/16 model architecture trained for 16B processed samples on LAION-CAT. Init: Initialization with random or SWAG-B/16 weights. ED: Embedding distillation. DD: Distribution distillation. LiT: Locked image tuning. FT: Fine-tuning. FT-delay: Locked image tuning for 50% followed by fine-tuning for the rest. CD: Our concept distillation using teacher-predicted objects and attributes.

Init	Method	SWAG (teacher)	IN	COCO T2I	COCO I2T	Flickr T2I	Flickr I2T
Random	Baseline	—	68.7	42.8	60.5	72.8	**89.7**
Random	ED	B/16	69.2	42.6	59.4	72.8	86.8
Random	DD	B/16	68.6	41.8	57.4	71.7	87.0
Random	CD (ours)	B/16	71.0	42.8	59.5	72.3	86.5
Random	CD (ours)	H/14	72.3	**43.4**	60.4	**73.8**	87.6
SWAG	LiT	—	**73.0**	32.5	50.6	60.8	79.6
SWAG	FT	—	71.2	43.1	60.3	73.1	87.7
SWAG	FT-delay	—	72.0	42.7	**60.7**	72.5	86.2

Notes:

- No training overhead as the predicted concepts are pre-computed.
- ED/DD is 60% slower with an 8% increase in GPU memory due to the need of running an additional copy of the vision tower.
- One could pre-compute embeddings for ED and DD as well (1.2TB), while our pre-computed predictions take only 32.6GB additional storage space when saving the top-10 predictions.
- Drawback of LiT/FT is that it requires the same architecture in the final setup, while our CD can be effortlessly combined with any architecture or training setup, by using stored predictions as metadata.

(3) Multimodal alignment with hard negatives



(3) Multimodal alignment with hard negatives

- InfoNCE loss [**Oord etal, 2018**]

$$\mathcal{L}_{\text{NCE}}(\mathbf{X}) = -\sum_{i=1}^{n}\left[\log \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{t}_i/\tau}}{\sum_j e^{\boldsymbol{x}_i^\top \boldsymbol{t}_j/\tau}} + \log \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{t}_i/\tau}}{\sum_j e^{\boldsymbol{x}_j^\top \boldsymbol{t}_i/\tau}}\right]$$

- We adapt [**Robinson etal, 2021**] loss with hard negative samples, for multi-modal training:

$$\mathcal{L}_{\text{HN-NCE}}(\mathbf{X}) = -\sum_{i=1}^{n}\log\left[\frac{e^{\boldsymbol{x}_i^\top \boldsymbol{t}_i/\tau}}{\alpha\cdot e^{\boldsymbol{x}_i^\top \boldsymbol{t}_i/\tau} + \sum_{j\neq i} e^{\boldsymbol{x}_i^\top \boldsymbol{t}_j/\tau} w^{i\to t}_{\boldsymbol{x}_i,\boldsymbol{t}_j}}\right] - \sum_{i=1}^{n}\log\left[\frac{e^{\boldsymbol{x}_i^\top \boldsymbol{t}_i/\tau}}{\alpha\cdot e^{\boldsymbol{x}_i^\top \boldsymbol{t}_i/\tau} + \sum_{j\neq i} e^{\boldsymbol{x}_j^\top \boldsymbol{t}_i/\tau} w^{t\to i}_{\boldsymbol{x}_j,\boldsymbol{t}_i}}\right]$$

$$\boxed{w^{i\to t}_{\boldsymbol{x}_i,\boldsymbol{t}_j} = \frac{(n-1)\cdot e^{\beta \boldsymbol{x}_i^\top \boldsymbol{t}_j/\tau}}{\sum_{k\neq i} e^{\beta \boldsymbol{x}_i^\top \boldsymbol{t}_k/\tau}}}$$

$$\boxed{w^{t\to i}_{\boldsymbol{x}_j,\boldsymbol{t}_i} = \frac{(n-1)\cdot e^{\beta \boldsymbol{x}_j^\top \boldsymbol{t}_i/\tau}}{\sum_{k\neq i} e^{\beta \boldsymbol{x}_k^\top \boldsymbol{t}_i/\tau}}}$$

- The weights w_β are designed such that difficult negative pairs are emphasized, and easier pairs are ignored. Furthermore, α rescales the normalization with the positive terms to account for the case when false negatives are present within the data.

(3) Multimodal alignment with hard negatives

LAION-CAT 438M dataset

Table 3. Evaluating effect of using hard negative contrastive loss. Evaluation performed on ViT-B/16 model architecture trained for 16B processed samples on LAION-CAT. CD: Our concept distillation using SWAG-H/14 predicted objects and attributes. HN: Our proposed hard negative contrastive loss.

#	Method		IN	COCO		Flickr	
	CD	HN		T2I	I2T	T2I	I2T
1			68.7	42.8	60.5	72.8	**89.7**
2	✓		**72.3**	43.4	60.4	**73.8**	87.6
3	✓	✓	72.0	**43.7**	**62.0**	73.2	89.5

PMD 63M public clean dataset

Table 4. Evaluating effect when pre-training on PMD using our approaches. Evaluation performed on ViT-B/32 and ViT-B/16 models trained for 4B processed samples. CD: Our concept distillation using SWAG-H/14 predicted objects (-O) and attributes (-A). HN: Our proposed hard negative contrastive loss.

Arch.	#	Method			IN	COCO		Flickr	
		CD-O	CD-A	HN		T2I	I2T	T2I	I2T
B/32	1				49.0	28.9	50.2	62.0	80.3
B/32	2	✓			57.8	32.2	54.0	65.6	85.7
B/32	3	✓	✓		59.7	34.4	55.7	68.3	87.8
B/32	4	✓	✓	✓	**62.4**	**37.3**	**60.4**	**71.8**	**89.9**
B/16	5				54.6	33.1	55.7	67.4	85.5
B/16	6	✓	✓		65.5	37.4	59.9	72.4	88.7
B/16	7	✓	✓	✓	**67.8**	**42.7**	**65.5**	**77.6**	**92.5**

Comparison with SOTA
DiHT - Distilled and Hard-negative Training

LAION-2B vs LAION-CAT 438M

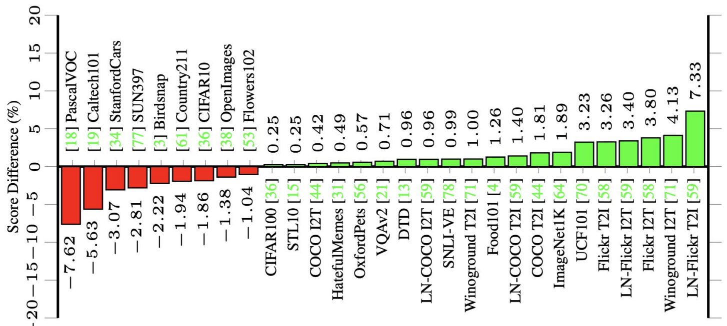


Figure 4. DiHT-B/16 trained on LAION-CAT with 438M samples *vs*. CLIP-B/16 trained on LAION-2B with 2B samples. Both models trained by us with 32B total processed samples.

PMD 63M public clean dataset

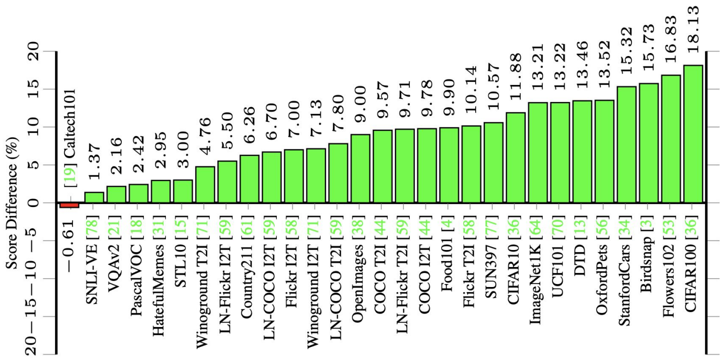


Figure 5. DiHT-B/16 *vs*. CLIP-B/16. Both models trained by us on PMD with 63M images and 4B total processed samples.

Comparison with SOTA
DiHT - Distilled and Hard-negative Training

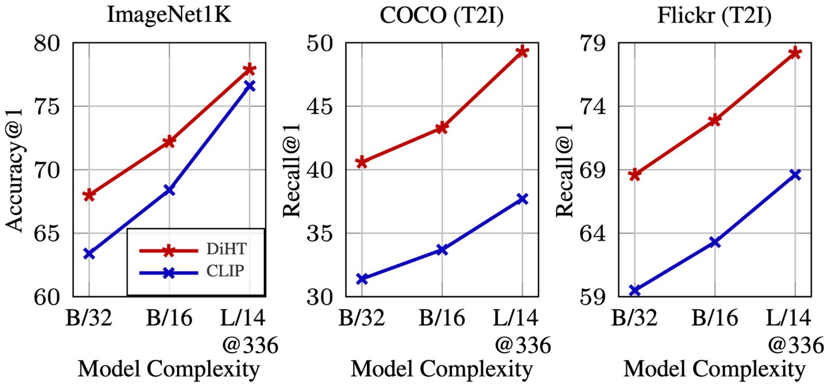


Figure 1. DiHT trained on 438M LAION-CAT samples *vs*. CLIP [61] trained on 400M OpenAI samples.

Method	px	#P	#D	#S	IN	COCO		Flickr	
						T2I	I2T	T2I	I2T
ViT-B/32									
CLIP [59]	224	151M	400M	12.8B	63.4	31.4	49.0	59.5	79.9
OpenCLIP [27]	224	151M	400M	12.8B	62.9	34.8	52.3	61.7	79.2
OpenCLIP [27]	224	151M	2.3B	34B	66.6	39.0	56.7	65.7	81.7
DiHT	224	151M	438M	16B	67.5	40.3	56.3	67.9	83.8
DiHT	224	151M	438M	32B	**68.0**	**40.6**	**59.3**	**68.6**	**84.4**
ViT-B/16									
CLIP [59]	224	150M	400M	12.8B	68.4	33.7	51.3	63.3	81.9
OpenCLIP [27]	224	150M	400M	12.8B	67.1	37.8	55.4	65.2	84.1
OpenCLIP [27]	240	150M	400M	12.8B	69.2	40.5	57.8	67.7	85.3
DiHT	224	150M	438M	16B	71.9	**43.7**	**62.0**	**73.2**	89.5
DiHT	224	150M	438M	32B	**72.2**	43.3	60.3	72.9	**89.8**
ViT-L/14									
CLIP [59]	224	428M	400M	12.8B	75.6	36.5	54.9	66.1	84.5
CLIP [59]	336	428M	400M	13.2B	76.6	37.7	57.1	68.6	86.6
OpenCLIP [27]	224	428M	400M	12.8B	72.8	42.1	60.1	70.4	86.8
OpenCLIP [27]	224	428M	2.3B	32B	75.2	46.2	64.3	75.4	**90.4**
DiHT	224	428M	438M	32B	75.9	47.7	**65.4**	76.8	90.0
DiHT	336	428M	438M	32.4B	**77.4**	**49.6**	65.1	**78.7**	90.2

Few-shot linear probing

- In practice, few-shot models perform significantly worse than zero-shot models in the low-data regime.
- Initializing with zero-shot classifiers, and learning with SGD using L2 penalty does not improve performance and the model simply ignores the supervision.

- We propose to ensure that the final weights do not drift much from the prompt using projected gradient descent (PGD).

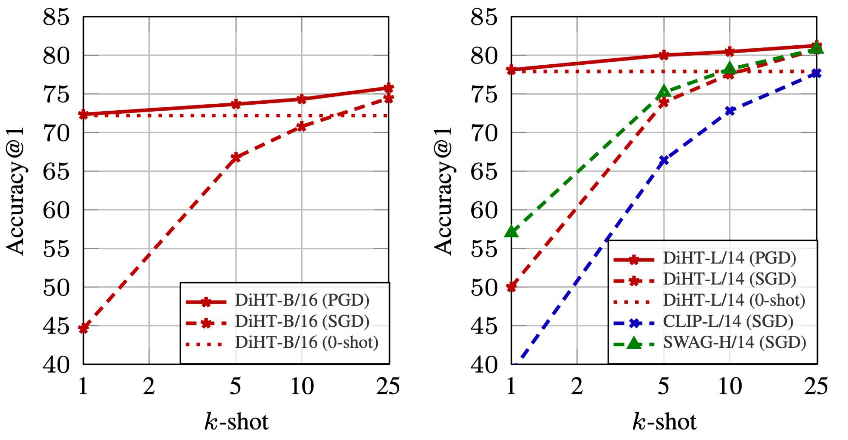


Figure 6. k-shot linear probing performance on ImageNet1K.