A. Qualitative Analysis

Fig. 8 presents more predicted examples including images with dogs and parks. It further shows that ImageNet based pre-training methods tend to map certain objects to a certain set of emotions. VisE, on the other hand, is able to predict correct emotions in these examples.

B. Supplementary Results and Discussion

Transfer learning on ImageNet Table 4 presents results and comparisons on ImageNet. We fine-tune VisE-250M with a ResNeXt-101 backbone on ImageNet, and compare the `val` accuracy scores with the same ResNeXt-101 model trained from scratch (IN-Sup). We also show the results of IG-940M-IN [55], which is pre-trained on 940 million images with 1.5K hashtags and fine-tuned on ImageNet using the same visual backbone. We see from Table 4 that representations learned from VisE-250M with engagement signals are transferable to ImageNet, outperforming the IN-Sup model by 0.88 (1.12%) measured by Top-1 accuracy. Note that engagement signals are relatively weak compared to the hashtags used in IG-940M-IN, which were selected to match with 1000 ImageNet synsets. Our goal here is to show features learned by VisE can be generalized to large-scale image classification tasks.

Images *vs*. engagement signals To disentangle the effect of training images and engagement signals, we also trained MoCo-v2 with the same 1.23 million social post data (VisE-1.2M(MoCo-v2)). Table 5 shows the linear evaluation results on UnbiasedEmotion, which shows the engagement signals, not the images, are beneficial for this dataset. We will include the full results in the final version.

Additional results Table 6 and 7 present full transfer learning results including performance on the `val` split and an additional metric for the Hateful Memes dataset. These two tables can be read in conjunction with the main figure and the backbone ablation studies in the main text. Note that we use in-house baselines instead of copying results from prior work for fair-comparison purposes. All the experiments are trained using the same grid search range, validation set, learning rate schedule, *etc*. We use validation accuracy and ROC AUC for Hateful Memes to select the best set of hyper-parameters. See Appendix C.3 for details.

Datasize calculation for contrastive learning methods In size ablation studies, we sort all pre-training methods by the training inputs size. We consider the negative input pairs for MoCo-v2 and CLIP as the *effective* training data size.

- **MoCo-v2** uses image pairs from ImageNet as inputs. The total class size is the total number of training data (1.28 million). The effective training data size is the number of image pairs used, which is (k + 1) × 1.28M = 83.9B, where $k = 65536$ is the number of images in the queue for MoCo-v2.
- **CLIP** uses a dataset with 400M image-text pairs. This approach considers the pair-wise similarity among image-text in a batch during training. Since the batch size is 32768, the total effective datasize is $400\text{M} / 32768 \times 32768 \times 32768 = 83.9\text{B}$.

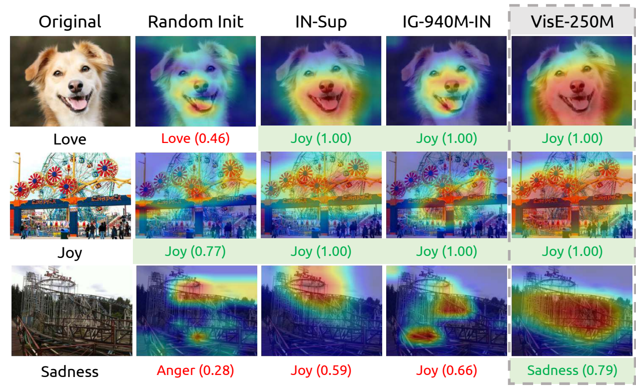

Figure 8. Qualitative results on UnbiasedEmotion dataset using ResNeXt-101 32×16d backbone.

Method	Top 1 Accuracy	Top 5 Accuracy
IN-Sup	78.78	94.12
VisE-250M	79.66 ↑0.88	94.62 ↑0.5
IG-940M-IN [55]	84.2	97.2

Table 4. Fine-tuned experiments on ImageNet with ResNeXt-101 32 × 16d backbone. Colored text with ↑ indicate the differences between VisE and IN-Sup.

Method	Val Accuracy	Test Accuracy
VisE-1.2M (MoCo-v2)	27.96 ± 2.91	27.80 ± 2.30
VisE-1.2M	44.67 ± 3.52 ↑16.71	45.74 ± 2.15 ↑17.94

Table 5. Linear evaluation experiments on UnbiasedEmotion with ResNet-50 using the same 1.23 million data.

C. Reproducibility Details

C.1. VisE Pre-training Setup

Optimization VisE models are trained on 32 GPUs across 4 machines with a batch size of 1920 images for the ResNet backbone and 1536 for the ResNeXt backbone. We use stochastic gradient descent with a momentum of 0.9 and a weight decay of 0.0001. The base learning rate is set according to $0.1/256 \times b$, where b is the batch size used for the particular model. The learning rate is warmed up linearly from 0 to the base learning rate during the first 5% of the whole iterations. The learning rate decay schedule is set differently for VisE-1.2M and VisE-250M. For models that use 1.23 million images, we follow common ImageNet pre-training settings. For models that are trained with 250 million images, the learning rate is reduced 10 times over approximately 10 epochs with the scaling factor of 0.5.

Training details We adopt standard image augmentation strategy during training (randomly resize crop to 224 × 224 and random horizontal flip). Since the dataset is not balanced, we

Backbone	Method	UnbiasedEmotion		Politics		Hateful Memes	
		Val Accuracy	Test Accuracy	Val Accuracy	Test Accuracy	ROC AUC	Accuracy
ResNet-50	Random Init	24.67 ± 2.78	23.80 ± 1.02	56.80	56.57	0.5335	51.64
	Uni-modality pre-training methods						
	IN-Sup	43.62 ± 2.31	44.36 ± 1.09	59.31	59.45	0.5691	53.16
	VQAGrid [35]	32.17 ± 1.22	33.57 ± 0.86	57.32	57.31	0.5517	53.2 ↑0.04
	Cross-modalities pre-training methods						
	VirTex [12]	40.59 ± 2.96	42.17 ± 1.14	58.46	58.44	0.5659	54.40 ↑1.24
	ICMLM$_{\text{att-fc}}$ [2]	23.81 ± 2.25	23.51 ± 1.52	58.27	58.41	0.5702 ↑0.0011	53.32 ↑0.16
	ICMLM$_{\text{tfm}}$ [2]	31.71 ± 2.02	31.87 ± 0.95	58.73	58.86	0.5631	53.24 ↑0.08
	Contrastive learning pre-training methods						
	MoCo-v2 [7]	26.31 ± 1.12	26.23 ± 1.20	58.14	58.30	0.5947 ↑0.0256	53.92 ↑0.76
	CLIP [66]	42.70 ± 3.02	45.41 ± 2.90 ↑1.05	56.65	56.42	**0.6147** ↑0.0456	**57.04** ↑3.88
	Ours						
	VisE-1.2M	44.67 ± 3.52 ↑1.05	45.74 ± 2.15 ↑1.38	59.15 ↓0.16	59.30 ↓0.15	0.6100 ↑0.0409	55.52 ↑2.36
	VisE-250M	**51.97** ± 4.08 ↑8.35	**53.05** ± 1.48 ↑8.69	**60.56** ↑1.25	**60.31** ↑0.86	0.5784 ↑0.0093	54.48 ↑1.32
ResNeXt-101 32 × 16d	Random Init	37.96 ± 3.77	38.43 ± 1.38	57.05	56.92	0.5466	53.64
	IN-Sup	63.09 ± 3.12	62.59 ± 1.99	59.24	59.42	0.5542	51.84
	IG-940M-IN [55]	55.86 ± 1.36	56.26 ± 1.32	60.98 ↑1.74	**61.15** ↑1.73	0.5482	52.28 ↑0.44
	VisE-1.2M (ours)	56.64 ± 2.49 ↓6.45	56.26 ± 1.05 ↓6.33	59.70 ↑0.46	59.89 ↑0.47	0.5621 ↑0.0079	54.24 ↑2.40
	VisE-250M (ours)	**69.61** ± 2.74 ↑6.51	**69.44** ± 1.20 ↑6.85	**61.08** ↑1.84	61.01 ↑1.59	**0.5795** ↑0.0253	**56.04** ↑4.20

Table 6. Linear evaluation experiments comparing VisE with other pre-training baselines. Colored text with ↑ and ↓ indicate the differences between VisE and IN-Sup with the same visual backbone. ↑ is also used if other methods yield better results than IN-Sup. In general, VisE outperforms the ImageNet supervised and hashtag-based weakly supervised pre-training methods.

Backbone	Method	UnbiasedEmotion		Politics		Hateful Memes	
		Val Accuracy	Test Accuracy	Val Accuracy	Test Accuracy	ROC AUC	Accuracy
ResNet-50	Random Init	39.01 ± 0.99	37.25 ± 2.12	58.28	58.31	0.5833	51.84
	Uni-modality pre-training methods						
	IN-Sup	69.87 ± 3.27	67.94 ± 3.18	63.87	63.64	0.6005	54.32
	VQAGrid [35]	42.17 ± 3.07	43.93 ± 1.56	58.31	58.1	0.5906	53.24
	Cross-modalities pre-training methods						
	VirTex [12]	72.24 ± 2.13 ↑2.37	73.61 ± 1.94 ↑5.67	63.24	63.06	0.5898	53.84
	ICMLM$_{\text{att-fc}}$ [2]	71.65 ± 2.31 ↑1.78	70.98 ± 2.01 ↑3.05	63.3	63.2	0.5846	53.52
	ICMLM$_{\text{tfm}}$ [2]	70.92 ± 1.65 ↑1.05	71.48 ± 1.78 ↑3.54	63.43	63.21	0.5842	53.40
	Contrastive learning pre-training methods						
	MoCo-v2 [7]	77.63 ± 1.78 ↑7.76	76.23 ± 1.88 ↑8.29	**66.24** ↑2.37	**66.37** ↑2.73	0.5884	52.48
	CLIP [66]	73.68 ± 0.93 ↑3.81	74.46 ± 1.21 ↑6.53	58.08	58.07	0.5470	53.48
	Ours						
	VisE-1.2M	73.82 ± 1.07 ↑3.95	74.20 ± 1.93 ↑6.26	64.69 ↑0.82	64.69 ↑1.05	**0.6070** ↑0.0039	**55.88** ↑0.96
	VisE-250M	**79.74** ± 1.54 ↑9.87	**78.89** ± 2.23 ↑10.95	65.83 ↑1.96	65.62 ↑1.98	0.6060 ↑0.0055	55.00 ↑0.68
ResNet-101	Random Init	40.20 ± 2.76	39.08 ± 1.72	58.18	58.07	0.5868	53.48
	IN-Sup	71.84 ± 2.72	72.43 ± 2.24	58.28	58.42	0.5939	**54**
	VisE-1.2M (ours)	**73.82** ± 0.77 ↑1.97	**74.52** ± 1.18 ↑2.10	**63.92** ↑5.64	**63.85** ↑5.43	**0.5958** ↑0.0019	52.96 ↓1.04
ResNeXt-101 32 × 16d	Random Init	40.20 ± 2.65	38.59 ± 0.91	58.26	58.39	0.5959	**54.68**
	IN-Sup	79.00 ± 2.33	77.92 ± 2.38	64.22	64.25	0.5903	52.92
	IG-940M-IN [55]	83.24 ± 1.68 ↑4.24	81.52 ± 1.76 ↑3.60	65.90 ↑1.68	65.58 ↑1.33	0.5951 ↑0.0048	54.28 ↑1.36
	VisE-1.2M (ours)	77.57 ± 2.43 ↓1.43	78.33 ± 1.39 ↑0.41	64.61 ↑0.39	64.44 ↑0.19	0.5976 ↑0.0073	54.40 ↑1.48
	VisE-250M (ours)	**84.08** ± 1.87 ↑5.08	**85.21** ± 1.24 ↑7.29	**67.61** ↑3.39	**67.64** ↑3.39	**0.5957** ↑0.0054	**54.96** ↑2.04

Table 7. Fine-tuning experiment comparing VisE with other pre-training baselines. Colored text with ↑ and ↓ indicate the differences with IN-Sup with the same visual backbone. ↑ is also used if other methods yield better results than IN-Sup. Similar to observations in Table 6, VisE can achieve better results compared to the ImageNet supervised and hashtag-based weakly supervised pre-training methods.

follow [50, 10] to stabilize the training processing by initializing the the bias for the last linear classification layer with $b = -\log\left((1-\pi)/\pi\right)$, where the prior probability π is set to 0.01. To obtain the pseudo-labels for the visual engagement signals, we set the number of clusters as 5000 and 128, for comments and raw reactions respectively.

Other details We spend around 9 hours to mine the data for pretraining with 3 server nodes (144 cpus). For the 1.23M data, the total word count for comments is 178M, the average $\pm_{\text{std}}$ number of comments per image is $20.25 \pm_{54.43}$, the average $\pm_{\text{std}}$ reactions count per image is $81.21 \pm_{601.4}$. We use Pytorch [63] to implement and train all the models on NVIDIA Tesla V100 GPUs.

Dataset	Task	# Classes	Train	Val	Test
Caltech-UCSD Birds-200-2011 [80]	Fine-grained bird species recognition	200	5994	5794	-
UnbiasedEmotion [62]	Image emotion recognition	6	2,131*	304*	610*
Politics [75]	Visual political bias prediction	2	607,306*	67,478*	75,148
Hateful Memes [40]	Hate speech detection in multimodal memes	2	8,500	500	-

Table 8. Specifications of the various target task dataset. Image number with * are the subset we randomly sampled since no publicly data splits are available. UnbiasedEmotion are randomly split 5 times.

Task	# GPUs	ResNet-50 (24M)		ResNet-101 (45M)		ResNeXt-101 (194M)	
		Per Iteration (second)	Total Time (minute)	Per Iteration (second)	Total Time (minute)	Per Iteration (second)	Total Time (minute)
UnbiasedEmotion	1	1.67	49.64	1.44	50.30	1.32	63.37
Politics	8	0.14	216.26	0.18	243.64	0.67	721.61
Hateful Memes	1	0.43	41.23	0.32	49.57	0.50	140.81
CUB-200-2011	1	0.74	294.01	-	-	0.91	1,304.01

Table 9. Average run time (per iteration and total) for fine-tuned experiments.

	Training Schedule		Method	Batch Size	Linear Evaluation				Fine-tuned			
					Base LR	WD	S_{lr}	S_{wd}	Base LR	WD	S_{lr}	S_{wd}
UnbiasedEmotion	Total epochs: 50 LR steps: (0, 10, 20, 30) LR decay: (1, 0.1, 0.01, 0.001)	ResNet-50	Random Init	128	0.025	{0.001, 0.01, 0.0001, 0.01, 0.01}	1.28 $\pm$ 0.38	0.36 $\pm$ 0.06	{0.0025, 0.025, 0.025, 0.0025, 0.0025 }	{0.0001, 0.001, 0.01, 0.01, 0.01}	4.39 $\pm$ 0.75	1.54 $\pm$ 0.54
			IN-Sup		0.025	{0.01, 0.0001, 0.01, 0.01, 0.01}	10.52 $\pm$ 0.65	0.38 $\pm$ 0.26	{0.0025, 0.0025, 0.0025, 0.0025, 0.025 }	{0.0001, 0.001, 0.001, 0.01, 0.0001}	7.35 $\pm$ 1.18	2.10 $\pm$ 0.50
			MoCo-v2		0.025	{0.01, 0.01, 0.0001, 0.01, 0.01}	3.06 $\pm$ 0.33	0.47 $\pm$ 0.18	0.0025	{0.0001, 0.01, 0.0001, 0.001, 0.0001}	3.94 $\pm$ 0.47	0.42 $\pm$ 0.28
			VQAGrid		0.025	{0.0001, 0.0001, 0.0001, 0.01, 0.01}	6.91 $\pm$ 0.57	0.51 $\pm$ 0.21	0.00025	{0.0001, 0.0001, 0.0001, 0.01, 0.01}	0.00 $\pm$ 0.00	0.64 $\pm$ 0.39
			VirTex		0.025	{0.001, 0.0001, 0.01, 0.0001, 0.01}	7.70 $\pm$ 0.45	0.41 $\pm$ 0.39	0.0025	{0.01, 0.0001, 0.001, 0.001, 0.001}	3.76 $\pm$ 0.64	0.81 $\pm$ 0.25
			$\text{ICMLM}_{\text{att-fc}}$		0.025	{0.001, 0.001, 0.01, 0.01, 0.0001}	2.20 $\pm$ 0.81	0.21 $\pm$ 0.07	0.025	{0.01, 0.0001, 0.0001, 0.001, 0.0001}	12.38 $\pm$ 1.33	0.76 $\pm$ 0.40
			$\text{ICMLM}_{\text{tfm}}$		0.025	{0.01, 0.01, 0.01, 0.01, 0.01}	6.08 $\pm$ 1.17	0.83 $\pm$ 0.35	0.025	{0.01, 0.0001, 0.01, 0.01, 0.0001}	6.81 $\pm$ 0.44	1.03 $\pm$ 0.46
			CLIP		0.025	{0.01, 0.01, 0.01, 0.01, 0.01}	9.88 $\pm$ 0.78	0.84 $\pm$ 0.29	2.5e-05	{0.01, 0.0001, 0.01, 0.0001, 0.001}	19.42 $\pm$ 0.79	4.36 $\pm$ 4.34
			VisE-1.2M		0.025	{0.001, 0.01, 0.01, 0.01, 0.01}	10.04 $\pm$ 0.75	0.81 $\pm$ 0.52	0.025	{0.01, 0.0001, 0.001, 0.001, 0.001}	3.36 $\pm$ 0.87	1.84 $\pm$ 0.83
			VisE-250M		0.025	0.01	14.35 $\pm$ 2.10	1.07 $\pm$ 0.24	0.0025	{0.001, 0.001, 0.001, 0.001, 0.01}	11.24 $\pm$ 2.84	1.77 $\pm$ 0.31
			VisE-123k	128	-	-	-	-	0.025	{0.01, 0.0001, 0.001, 0.001, 0.0001}	3.62 $\pm$ 0.84	0.68 $\pm$ 0.39
			VisE-308k		-	-	-	-	0.025	{0.001, 0.001, 0.001, 0.01, 0.0001}	3.21 $\pm$ 0.87	1.07 $\pm$ 0.63
			VisE-615k		-	-	-	-	{0.0025, 0.0025, 0.025, 0.025, 0.025}	{0.0001, 0.0001, 0.0001, 0.001, 0.01}	2.98 $\pm$ 0.97	1.08 $\pm$ 0.64
			VisE-1.2M-$\mathcal{C}$		-	-	-	-	{0.0025, 0.0025, 0.0025, 0.025, 0.025 }	{0.001, 0.0001, 0.0001, 0.001, 0.01}	3.00 $\pm$ 0.72	1.09 $\pm$ 0.79
			VisE-$\mathcal{R}$		-	-	-	-	{0.025, 0.0025, 0.025, 0.0025, 0.025 }	{0.001, 0.001, 0.001, 0.01, 0.01}	2.18 $\pm$ 0.63	0.94 $\pm$ 0.78
		ResNet-101	Random Init	64	-	-	-	-	{0.0025, 0.00025, 0.00025, 0.00025, 0.00025}	{0.01, 0.0001, 0.001, 0.0001, 0.0001}	5.79 $\pm$ 0.67	2.00 $\pm$ 1.08
			IN-Sup		-	-	-	-	0.0025	{0.01, 0.0001, 0.001, 0.01, 0.0001}	7.76 $\pm$ 2.69	1.23 $\pm$ 0.51
			VisE-1.2M		-	-	-	-	0.025	{0.01, 0.0001, 0.0001, 0.0001, 0.0001}	3.16 $\pm$ 0.74	1.04 $\pm$ 0.38
		ResNeXt-101	Random Init	32	0.025	0.0001	2.33 $\pm$ 1.12	0.30 $\pm$ 0.13	0.025	{0.001, 0.0001, 0.01, 0.001, 0.001}	2.74 $\pm$ 0.88	0.93 $\pm$ 0.36
			IN-Sup		0.025	{0.0001, 0.0001, 0.0001, 0.01, 0.0001}	9.92 $\pm$ 1.04	0.12 $\pm$ 0.12	0.025	{0.01, 0.0001, 0.001, 0.01, 0.001}	12.79 $\pm$ 1.48	1.37 $\pm$ 0.70
			IG-940M-IN		0.025	{0.0001, 0.0001, 0.01, 0.0001, 0.0001}	10.15 $\pm$ 1.42	0.19 $\pm$ 0.12	0.025	{0.001, 0.0001, 0.001, 0.001, 0.001}	15.21 $\pm$ 5.96	1.25 $\pm$ 0.43
			VisE-1.2M		0.025	{0.01, 0.001, 0.01, 0.01, 0.01}	13.86 $\pm$ 0.80	0.32 $\pm$ 0.20	0.0025	{0.0001, 0.001, 0.0001, 0.01, 0.001}	19.07 $\pm$ 8.31	1.60 $\pm$ 0.29
			VisE-250M		0.025	{0.0001, 0.01, 0.0001, 0.0001, 0.0001}	11.29 $\pm$ 0.79	0.12 $\pm$ 0.15	0.0025	{0.01, 0.001, 0.0001, 0.01, 0.0001}	13.73 $\pm$ 0.95	0.31 $\pm$ 0.22

Table 10. Hyperparameter configurations for best-performing UnbiasedEmotion models for five random split. Single number are displayed if the configurations are the same across all five experiments.

C.2. Other Pre-training Methods

We use the publicly available pre-trained models for other compared baseline methods[4] except for ImageNet pretraining with ResNeXt-101 backbone. We train that model with 100 epochs with learning rate decay schedule of (30, 60, 90) and scaling factor of 0.1. Note that the pre-trained model for CLIP adopts a modified ResNet-50 architecture. See [66] for details.

[4]Links for the publicly available pre-trained models: IG-940M-IN, MoCo-v2, VQAGrid, VirTex, ICMLM, CLIP.

C.3. Downstreaming Tasks Setup

Tasks summary The statistics of these tasks and the associated datasets are listed in Table 8.

	Training Schedule	Backbone	Method	Batch Size	Linear Evaluation				Fine-tuned			
					Base LR	WD	S_{lr}	S_{wd}	Base LR	WD	S_{lr}	S_{wd}
Politics	Total epochs: 25 LR steps: (0, 10, 20) LR decay: (1, 0.1, 0.01)	ResNet-50	Random Init	192	0.0025	0.01	2.37	1.49	0.025	0.0001	0.66	0.80
			IN-Sup		0.025	0.001	2.72	0.07	0.0025	0.001	1.81	0.82
			MoCo-v2		0.025	0.0001	1.83	0.59	0.025	0.0001	0.75	3.46
			VQAGrid		0.0025	0.01	2.02	0.36	0.00025	0.001	0.00	0.43
			VirTex		0.025	0.001	2.58	0.44	0.025	0.0001	0.19	2.14
			$\text{ICMLM}_{\text{att-fc}}$		0.025	0.001	2.37	0.39	0.025	0.0001	1.06	2.05
			$\text{ICMLM}_{\text{tfm}}$		0.025	0.001	2.90	0.31	0.025	0.0001	0.69	2.48
			CLIP		0.025	0.001	1.54	0.05	0.000025	0.001	0.45	0.24
			VisE-1.2M		0.025	0.001	2.60	0.27	0.025	0.0001	0.61	2.23
			VisE-250M		0.025	0.001	2.69	0.34	0.00025	0.01	3.83	0.15
			VisE-123k	192	-	-	-	-	0.025	0.0001	0.39	0.11
			VisE-308k		-	-	-	-	0.025	0.0001	1.01	2.41
			VisE-615k		-	-	-	-	0.025	0.0001	0.72	1.71
			VisE-1.2M-$\mathcal{C}$		-	-	-	-	0.025	0.0001	0.42	2.54
			VisE-$\mathcal{R}$		-	-	-	-	0.025	0.0001	0.36	2.17
		ResNet-101	Random Init	192	-	-	-	-	0.025	0.001	0.48	0.69
			IN-Sup		-	-	-	-	0.025	0.0001	0.61	0.77
			VisE-1.2M		-	-	-	-	0.025	0.0001	0.49	2.16
		ResNeXt-101	Random Init	192	0.0025	0.001	0.07	0.05	0.025	0.0001	0.61	0.80
			IN-Sup		0.0025	0.0001	0.08	0.06	0.025	0.0001	1.26	1.76
			IG-940M-IN		0.0025	0.001	3.36	0.01	0.0025	0.0001	2.63	1.37
			VisE-1.2M		0.025	0.0001	0.59	0.48	0.025	0.0001	0.80	2.74
			VisE-250M		0.025	0.001	0.64	0.53	0.025	0.0001	0.16	2.03

Table 11. Hyperparameter configurations for best-performing Politics models. "Batch Size" presents the total mini batch size across 8GPUs. For fine-tuned settings, some learning processes are stopped early.

	Training Schedule	Backbone	Method	Batch Size	Linear Evaluation				Fine-tuned			
					Base LR	WD	S_{lr}	S_{wd}	Base LR	WD	S_{lr}	S_{wd}
Hateful Memes	Total epochs: 30 LR steps: (0, 20) LR decay: (1, 0.5)	ResNet-50	Random Init	64	0.025	0.01	0.0184	0.0022	0.025	0.01	0.0322	0.0014
			IN-Sup		0.025	0.01	0.0184	0.0006	0.025	0.0001	0.0346	0.0064
			MoCo-v2		0.025	0.01	0.0364	0.0009	0.025	0.001	0.0265	0.0022
			VQAGrid		0.025	0.01	0.0265	0.001	0.025	0.0001	0.0442	0.0426
			VirTex		0.025	0.01	0.0194	0.0018	0.025	0.001	0.0301	0.0011
			$\text{ICMLM}_{\text{att-fc}}$		0.025	0.01	0.018	0.0002	0.025	0.001	0.0353	0.0011
			$\text{ICMLM}_{\text{tfm}}$		0.025	0.01	0.0217	0.002	0.025	0.01	0.0395	0.0024
			CLIP		0.025	0.01	0.0583	0.0007	0.00025	0.01	0	0.0021
			VisE-1.2M		0.025	0.01	0.0453	0.0015	0.025	0.01	0.0403	0.0061
			VisE-250M		0.025	0.01	0.0191	0.0018	0.025	0.0001	0.0284	0.0063
			VisE-123k	64	-	-	-	-	0.025	0.01	0.0452	0.0033
			VisE-308k		-	-	-	-	0.025	0.01	0.0324	0.0014
			VisE-615k		-	-	-	-	0.025	0.0001	0.0341	0.0074
			VisE-1.2M-$\mathcal{C}$		-	-	-	-	0.025	0.001	0.029	0.0033
			VisE-$\mathcal{R}$		-	-	-	-	0.025	0.001	0.0318	0.0043
		ResNet-101	Random Init	32	-	-	-	-	0.025	0.01	0.0404	0.0062
			IN-Sup		-	-	-	-	0.025	0.01	0.0378	0.0053
			VisE-1.2M		-	-	-	-	0.025	0.01	0.0368	0.0088
		ResNeXt-101	Random Init	16	0.00025	0.0001	0.0046	0	0.025	0.0001	0.0365	0.0015
			IN-Sup		0.025	0.01	0.02	0.0006	0.025	0.001	0.0348	0.009
			IG-940M-IN		0.025	0.01	0.0308	0.001	0.025	0.01	0.032	0.0018
			VisE-1.2M		0.025	0.01	0.0273	0.0016	0.025	0.01	0.0404	0.0021
			VisE-250M		0.025	0.01	0.0161	0.0018	0.025	0.01	0.0324	0.0051

Table 12. Hyperparameter configurations for best-performing Hateful Memes models. The text encoder is used as a feature extractor in these experiments.

Implementation Similar to the pre-training models, we use Pytorch and NVIDIA Tesla V100 16GB GPUs for the transfer learning experiments. Table 9 summarizes other implementation details including average runtime. The same data augmentation are employed as the pretraining stage. To encode raw text of the multi-modal experiments, we use RoBERTa base from fairseq [61][5].

Optimization and training details We use stochastic gradient descent with 0.9 momentum for image only models and Adam optimization with decoupled weight decay [53] for multi-modal experiments. Following [55], we conduct a coarse grid search to find the learning rate and weight decay values using `val` split. The learning rate is set as Base LR/256 × batchsize, where Base LR is chosen from $\{0.025, 0.0025, 0.00025\}$. For pre-training method CLIP, we expand the search to $\{0.025, 0.0025, 0.00025, 0.000025, 0.0000025\}$. The bound for weight decay is: $\{0.01, 0.001, 0.0001\}$. We also report the model performance sensitivity to learning rate (S_{lr}) and weight decay (S_{wd}) values. S_{lr} is defined as the standard deviation of the model performance across the range of learning rate considered given the optimal weight decay value. Similarly, S_{wd} is the standard deviation across the range of weight decay values given the optimal learning rate. Tables 10-16 show the training details and hyperparameter configurations of all the experiments in the main text.

[5] Link for the publicly available pre-trained RoBERTa-base model

	Training Schedule	Backbone	Method	Batch Size	Linear Evaluation				Fine-tuned			
					Base LR	WD	S_{lr}	S_{wd}	Base LR	WD	S_{lr}	S_{wd}
CUB-200-2011	Total epochs: 300 LR steps: (0, 100, 200) LR decay: (1, 0.1, 0.01)	ResNet-50	Random Init	128	0.025	0.001	1.17	0.20	0.025	0.01	23.12	14.49
			IN-Sup		0.025	0.001	25.25	0.62	0.025	0.01	10.14	0.99
			VisE-1.2M		0.025	0.0001	4.04	0.75	0.025	0.001	10.41	1.31
			VisE-250M		0.025	0.0001	3.87	0.94	0.025	0.001	4.71	0.80
		ResNeXt-101	Random Init	32	0.025	0.0001	1.86	0.91	0.025	0.01	27.50	8.04
			IN-Sup		0.025	0.0001	8.83	0.06	0.025	0.0001	2.97	3.23
			IG-940M-IN		0.025	0.0001	23.98	0.72	0.0025	0.001	1.13	0.00
			VisE-1.2M		0.025	0.001	3.28	0.96	0.025	0.001	29.39	0.60
			VisE-250M		0.025	0.0001	3.19	1.22	0.025	0.001	30.14	2.42

Table 13. Hyperparameter configurations for best-performing CUB-200-2011 models.

	Training Schedule	Backbone	Method	Batch Size	Image + Text (Fine-tuned)			
					Base LR	WD	S_{lr}	S_{wd}
Hateful Memes	Total epochs: 30 LR steps: (0, 20) LR decay: (1, 0.5)	ResNet-50	Random Init	64	0.00025	0.001	0.04	0.0099
			IN-Sup		0.00025	0.001	0.0613	0.0052
			MoCo-v2		0.00025	0.01	0.0668	0.0005
			VQAGrid		0.00025	0.001	0.0497	0.0067
			VirTex		0.00025	0.001	0.0704	0.037
			$\text{ICMLM}_{\text{att-fc}}$		0.00025	0.01	0.0684	0.0029
			$\text{ICMLM}_{\text{tfm}}$		0.00025	0.0001	0.0713	0.0073
			CLIP		0.00025	0.001	0	0.0026
			VisE-1.2M		0.00025	0.0001	0.0662	0.0039
			VisE-250M		0.00025	0.01	0.0697	0.0045
		ResNeXt-101	IG-940M-IN	16	0.00025	0.001	0.0628	0.0022
			VisE-250M		0.00025	0.01	0.0716	0.0052

Table 14. Hyperparameter configurations for best-performing Hateful Memes models: Image + Text (Fine-tuned).

	Training Schedule	Backbone	Method	Batch Size	Image Only (Fine-tuned)			
					Base LR	WD	S_{lr}	S_{wd}
Hateful Memes	Total epochs: 30 LR steps: (0, 20) LR decay: (1, 0.5)	ResNet-50	Random Init	64	0.0025	0.0001	0.0058	0.0043
			IN-Sup		0.0025	0.001	0.0206	0.0099
			MoCo-v2		0.025	0.01	0.0117	0.0015
			VQAGrid		0.00025	0.001	0	0.0154
			VirTex		0.025	0.0001	0.007	0.0171
			$\text{ICMLM}_{\text{att-fc}}$		0.025	0.01	0.01	0.0186
			$\text{ICMLM}_{\text{tfm}}$		0.025	0.001	0.0146	0.0224
			CLIP		0.00025	0.001	0	0
			VisE-1.2M		0.025	0.0001	0.0168	0.0044
			VisE-250M		0.0025	0.01	0.0185	0.0146
		ResNeXt-101	IG-940M-IN	16	0.00025	0.001	0.0146	0.0023
			VisE-250M		0.0025	0.001	0.0161	0.0122

Table 15. Hyperparameter configurations for best-performing Hateful Memes models: Image Only (Fine-tuned).

	Training Schedule	Backbone	Method	Batch Size	Image Only (Linear)			
					Base LR	WD	S_{lr}	S_{wd}
Hateful Memes	Total epochs: 30 LR steps: (0, 20) LR decay: (1, 0.5)	ResNet-50	Random Init	64	0.0025	0.0001	0.0075	0
			IN-Sup		0.025	0.01	0.0055	0.0002
			MoCo-v2		0.025	0.0001	0.0157	0.0004
			VQAGrid		0.00025	0.0001	0.0125	0.0004
			VirTex		0.025	0.01	0.0078	0.0007
			$\text{ICMLM}_{\text{att-fc}}$		0.025	0.001	0.006	0.0001
			$\text{ICMLM}_{\text{tfm}}$		0.025	0.01	0.0179	0.0001
			CLIP		0.025	0.01	0.043	0.0004
			VisE-1.2M		0.0025	0.0001	0.0089	0
			VisE-250M		0.025	0.01	0.0291	0.0009
		ResNeXt-101	IG-940M-IN	16	0.025	0.01	0.0092	0.0004
			VisE-250M		0.025	0.01	0.0299	0.0018

Table 16. Hyperparameter configurations for best-performing Hateful Memes models: Image Only (Linear).