南京航空航天大學

Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室

MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Exploring data bias in the VLM model

Dataset classification task

Larger, more diversified → less biased dataset

To some extent, but not entirely.

- The “Name of that dataset” experiment[1][2]
  - **Large** in scale
  - **General** and **diversified**
  - Purpose of **pre-training**

dataset	description
YFCC [49]	100M Flickr images
CC [4]	12M Internet image-text pairs
DataComp [15]	1B image-text pairs from Common Crawl
WIT [45]	11.5M Wikipedia images-text pairs
LAION [42]	2B image-text pairs from Common Crawl
ImageNet [8]	14M images from search engines

[1] Antonio Torralba & Alexei A. Efros. Unbiased look at dataset bias. CVPR 2011.
[2] Liu, Zhuang, Kaiming He, Meta AI Research. A decade’s battle on dataset bias: Are we there yet? ICLR 2025.

A Decade's Battle on Dataset Bias: Are We There Yet?

Zhuang Liu Kaiming He*

Meta AI Research, FAIR

Observation

➢ **High accuracy is observed across dataset combinations**

- **High accuracy**: All combinations >62%, 16 of 20 combinations >80%.
- **Highest combinations** : YFCC, CC, and ImageNet achieved 92.7%.
- **More datasets setting**: 6-dataset combination reached 69.2%.
- **Random baseline**: 33.3%.

YFCC	CC	DataComp	WIT	LAION	ImageNet	accuracy
✓	✓	✓				84.7
✓	✓		✓			83.9
✓	✓			✓		85.0
✓	✓				✓	92.7
✓		✓	✓			85.8
✓		✓		✓		72.1
✓		✓			✓	90.2
✓			✓	✓		86.6
✓			✓		✓	86.7
✓				✓	✓	91.9
	✓	✓	✓			83.6
	✓	✓		✓		62.8
	✓	✓			✓	82.8
	✓		✓	✓		84.3
	✓		✓		✓	91.3
	✓			✓	✓	84.1
		✓	✓	✓		71.5
		✓	✓		✓	88.9
		✓		✓	✓	68.2
			✓	✓	✓	90.7
✓	✓	✓				84.7
✓	✓	✓	✓			79.1
✓	✓	✓	✓	✓		67.4
✓	✓	✓	✓	✓	✓	69.2

Experiment on ConvNeXt-T

[2] Liu, Zhuang, Kaiming He, Meta AI Research. A decade's battle on dataset bias: Are we there yet? ICLR 2025.

Observation

- High accuracy is observed across dataset combinations
- **High accuracy is observed across model architectures**

- **High accuracy**: 4 of 5 models >80%, which AlexNet achieves 77.8%.
- **Architecture independent**: Networks capture biases regardless of architecture.
- **Inherent ability**: Bias capture is a core property of deep networks.

model	accuracy
AlexNet	77.8
VGG-16	83.5
ResNet-50	83.8
ViT-S	82.4
ConvNeXt-T	84.7

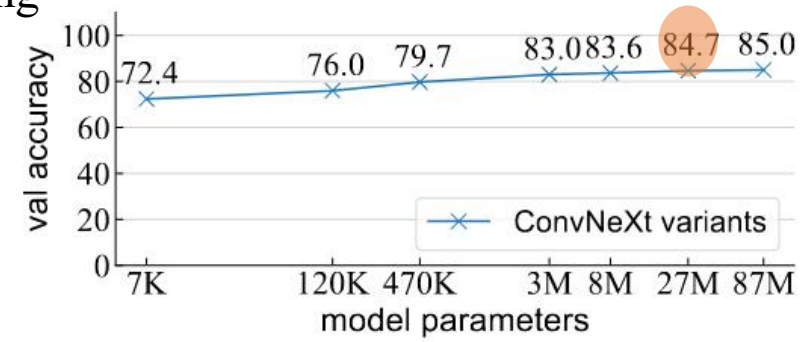
Experiment on the YCD combina

[2] Liu, Zhuang, Kaiming He, Meta AI Research. A decade's battle on dataset bias: Are we there yet? ICLR 2025.

Observation

- High accuracy is observed across dataset combinations
- High accuracy is observed across model architectures
- **High accuracy is observed across different model sizes**

  - **Small models still effective:** 7K-parameter ConvNeXt (3/10000 of ResNet-50) achieved 72.4% accuracy,
  - **Larger models perform better**: Accuracy improves with model size, but with diminishing returns
  - **No overfitting**: the presence of generalizable



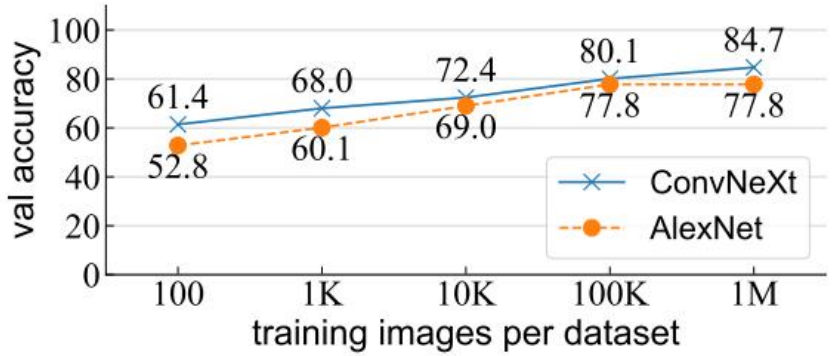
ConvNeXt-T Equivalent to parametric 27M model

[2] Liu, Zhuang, Kaiming He, Meta AI Research. A decade's battle on dataset bias: Are we there yet? ICLR 2025.

Observation

- High accuracy is observed across dataset combinations
- High accuracy is observed across model architectures
- High accuracy is observed across different model sizes
- **Dataset classification accuracy benefits from more training data**

- **More data improves validation accuracy**
- **Consistency across models:** observed in both modern ConvNeXt and classic AlexNet.
- **Generalization ability:** Models learn generalizable semantic patterns rather than simply memorizing training data.



Similar the behavior in semantic classification tasks

[2] Liu, Zhuang, Kaiming He, Meta AI Research. A decade's battle on dataset bias: Are we there yet? ICLR 2025.

Observation

- High accuracy is observed across dataset combinations
- High accuracy is observed across model architectures
- High accuracy is observed across different model sizes
- Dataset classification accuracy benefits from more training data
- **Dataset classification accuracy benefits from data augmentation**

  - **Effect of data augmentation:** Similar to increasing dataset size, it makes memorizing training images harder
  - **Improves accuracy:** Stronger data augmentation
  - **Generalization:** Models learn generalizable patterns

augmentation / training images per dataset	10K	100K	1M
no aug	43.2	71.9	76.8
w/ RandCrop	66.1	74.5	84.2
w/ RandCrop, RandAug	70.2	78.0	85.0
w/ RandCrop, RandAug, MixUp/CutMix	72.4	80.1	84.7

Similar the behavior in semantic classification tasks

[2] Liu, Zhuang, Kaiming He, Meta AI Research. A decade's battle on dataset bias: Are we there yet? ICLR 2025.

Exploration

- **Low-level Signatures**

corruption (on train+val)	accuracy
none	84.7
color jittering (strength: 1.0)	81.1
color jittering (strength: 2.0)	80.2
Gaussian noise (std: 0.2)	77.3
Gaussian noise (std: 0.3)	75.1
Gaussian blur (radius: 3)	80.9
Gaussian blur (radius: 5)	78.1
low resolution (64×64)	78.4
low resolution (32×32)	68.4

- **Pseudo-dataset classification task**

imgs per set	w/o aug	w/ aug
100	100.0	100.0
1K	100.0	100.0
10K	100.0	fail
100K	fail	fail

Table 6: *Training* accuracy on a pseudo-dataset classification task. Here we create 3 pseudo-datasets, all of which are sampled without replacement from the same source dataset (YFCC). This *training* task becomes more difficult for the network to solve if given more training images and/or stronger data augmentation. Validation accuracy is ~33% as no transferrable pattern is learned.

[2] Liu, Zhuang, Kaiming He, Meta AI Research. A decade's battle on dataset bias: Are we there yet? ICLR 2025.

Exploration

- **Pre-training task experiments**

case	accuracy
fully-supervised	82.9
linear probing w/	
MAE trained on IN-1K	76.2
MAE trained on YCD	78.4

Table 7: Self-supervised pre-training, followed by linear probing, achieves high accuracy for dataset classification. Here, we study MAE [19] as our self-supervised pre-training baseline, which uses ViT-B as the backbone. The fully-supervised baseline for dataset classification is with the same ViT-B architecture (82.9%). Results are on the YCD combination.

case	transfer acc
random weights	6.7
Y+C+D	27.7
Y+C+D+W	34.2
Y+C+D+W+L	34.2
Y+C+D+W+L+I	34.8
MAE [19]	68.0
MoCo v3 [5]	76.7

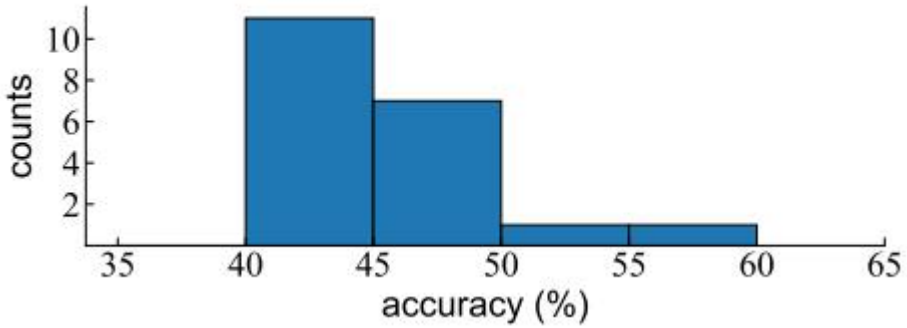
Table 8: Features learned by classifying datasets can achieve nontrivial results under the linear probing protocol. Transfer learning (linear probing) accuracy is reported on ImageNet-1K, using ViT-B as the backbone in all entries. The acronyms follow the first letter of each dataset in Table 2.

- **Transfer learning similarity**: Similar to semantic classification tasks.
- **Semantic features relevance**: Dataset biases align with features useful for image classification.

[2] Liu, Zhuang, Kaiming He, Meta AI Research. A decade's battle on dataset bias: Are we there yet? ICLR 2025.

Exploration

- **Human performance**



  - YFCC: people (6), scenery (3), natural lighting, plants, lifestyle (2), real-world, sport, wedding, high resolution (2), darker, most specific, most new, cluttered;
  - CC: cartoon (2), animated, clothing sample, product, logo, concept, explanatory texts, geography, furniture, animals, low resolution, colorful, brighter, daily images, local images, single person, realistic, clean background;
  - DataComp: white background (3), white space, transparent background, cleaner background, single item (2), product (2), merchandise, logo-style, product showcase, text (2), lots of words, artistic words, ads, stickers, animated pictures (2), screenshots, close-up shot, single person, people, non-realistic icons, cartoon, retro;

[2] Liu, Zhuang, Kaiming He, Meta AI Research. A decade's battle on dataset bias: Are we there yet? ICLR 2025.

Summaries

- **Key finding:** Modern neural networks easily capture dataset biases, showing robustness across different models and settings.
- **Nature of biases:** The biases captured by neural networks may include **generalizable and transferable patterns**, but their exact nature remains unclear and often imperceptible to humans.
- **Future direction:** Further investigation into this issue could help build datasets with fewer biases, improving model performance and fairness.

[2] Liu, Zhuang, Kaiming He, Meta AI Research. A decade's battle on dataset bias: Are we there yet? ICLR 2025.

Category

➢ Select bias[3]

It occurs when individuals or groups in a study **differ systematically from** the population of interest leading to a systematic error in an association or outcome.

Dataset	Bias Description	Impact
Caltech101 (2004)	Car images are mostly **side views**, showing angle bias.	Limits the model's ability to recognize cars from other perspectives.
ImageNet (2009)	Contains more **racing cars**, showing category preference.	Affects the model's generalization to regular cars and other categories.
Adience (2014)	**Dark-skinned females only make up 7.4%**, showing severe imbalance.	Results in lower classification accuracy for dark-skinned females.
IJB-A (2015)	**Dark-skinned females only make up 4.4%**, with lighter-skinned males dominating.	Leads to significantly higher error rates for dark-skinned individuals.
Berkeley Driving Dataset (2020)	Driving scenes are from only **four U.S. cities**.	Poor performance in other cities with different visual characteristics.

[3] Fabbrizzi S & Papadopoulos S. Efros. A survey on bias in visual datasets. COMPUT VIS IMAGE UND 2022.

Instance

➢ **Select bias**

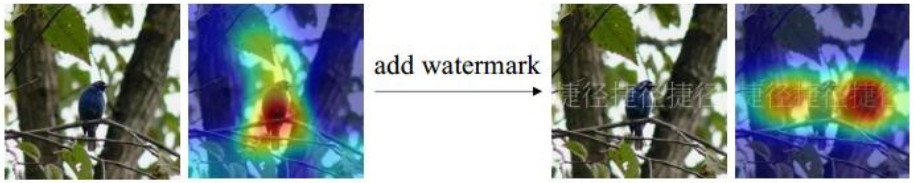
background shortcut

co-occurring object shortcut

UrbanCars

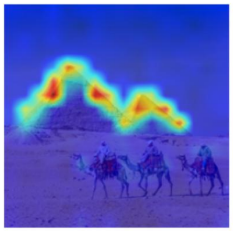
Carton class in ImageNet-1k

ImageNet-W

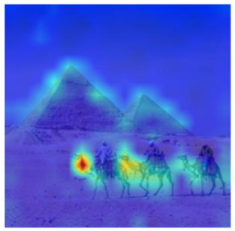


prediction: indigo bunting

prediction: carton



"A photo of a pyramid"



"A photo of a camel"

WIT training for CLIP

LAION training for CLIP

Category

➢ Framing bias

Framing bias as any association or disparities that can be used to **convey different messages** and/or that can be traced back to the way in which the visual content has been composed.

Example	Description	Impact
Obesity portrayal bias	59% of obese individuals depicted as **headless**.	Causes stigmatization and dehumanization.
Facial prominence bias	Men's faces more prominent in media and art.	Reflects gender-based visual bias.
Search engine bias	Women shown as **unprofessional or sexualized**.	Spreads stereotypes and misinformation.

Category

➢ Label bias

Semantic categories are often **not well defined** and different annotators may assign different labels to objects of the same type.

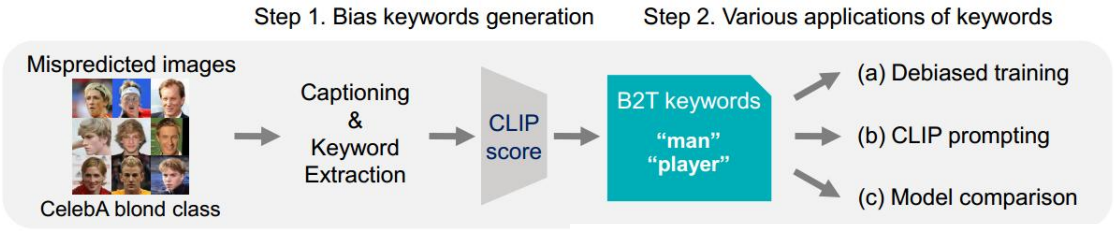
Example	Description	Impact
Semantic category bias	Labels like "grass" vs. "lawn" highlight inconsistencies in annotation.	Leads to inconsistent data and learning issues.
Race and skin color bias	Race classification relies on subjective judgment; skin tone suggested as an alternative.	Fails to capture diversity and introduces stereotypes.
Facial aesthetics bias	Subjective biases of annotators affect facial beauty datasets.	Reduces model generalization and fairness.

Discovering and Mitigating Visual Biases through Keyword Explanation

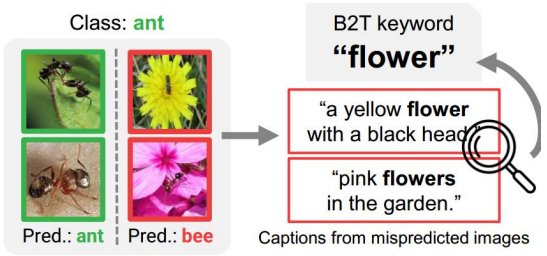
Younghyun Kim*[1] Sangwoo Mo*[2] Minkyu Kim[3] Kyungmin Lee[1] Jaeho Lee[4] Jinwoo Shin[1]
[1]KAIST [2]University of Michigan [3]KRAFTON [4]POSTECH
younghyun.kim@kaist.ac.kr swmo@umich.edu

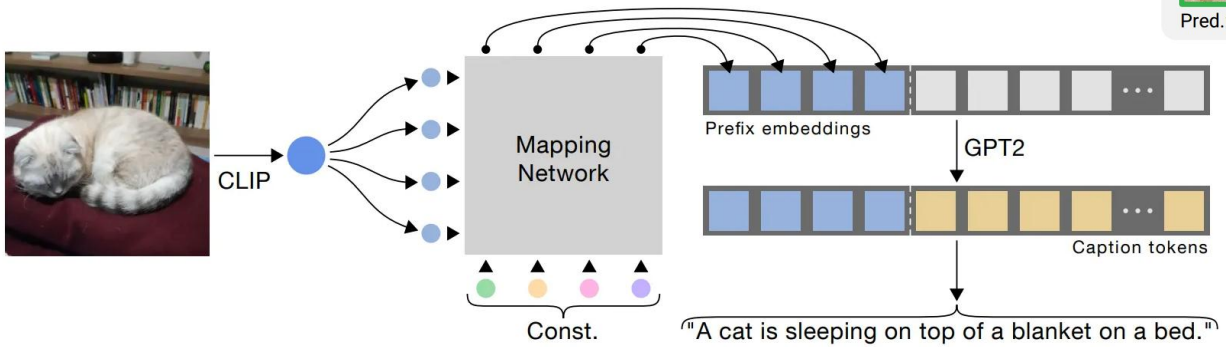
Aspect 1 — Keyword

➢ **Method**: B2T: Bias-to-text



Step1: use ClipCap as our default **captioning model**





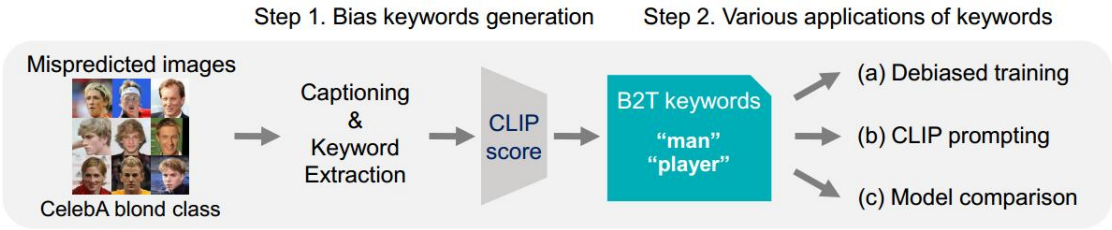
Step2: apply the YAKE algorithm to **extract keywords**

Text Preprocessing (Segmentation) --> Feature Extraction -->
Individual Word Weight Calculation --> Candidate Keyword Generation

[4] Kim Y et al. Discovering and Mitigating Visual Biases through Keyword Explanation. CVPR, 2024.

Aspect 1 — Keyword

➢ **Method**: B2T: Bias-to-text



Step3: **verify** that keywords represent bias by CLIP score

To **measures the similarity** between the keywords and the incorrectly predicted images

$$s_{\text{CLIP}}(a;\mathcal{D}) := \text{sim}(a,\mathcal{D}_{\text{wrong}}) - \text{sim}(a,\mathcal{D}_{\text{correct}})$$

$$\text{sim}(a,\mathcal{D}) := \frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} f_{\text{image}}(x)\, f_{\text{text}}(a).$$

[4] Kim Y et al. Discovering and Mitigating Visual Biases through Keyword Explanation. CVPR, 2024.

Aspect 1 — Keyword

➢ **Application: CLIP prompting**

Table 8. Prompt designs for debiaisng zero-shot classifiers.

Dataset	Dataset-wise Template	Class Name
CelebA	• [class name] • [class name] man • [class name] player • [class name] person • [class name] artist • [class name] comedy • [class name] film • [class name] actor • [class name] face	1. Blond • blond hair • celebrity of blond hair 2. Non blond • non blond hair • celebrity of non blond hair
Waterbirds	• [class name] • [class name] on the forest • [class name] with woods • [class name] on a tree • [class name] on a branch • [class name] in the forest • [class name] on the tree • [class name] on the ocean • [class name] on a beach • [class name] on the lake • [class name] with a surfer • [class name] on the water • [class name] on a boat • [class name] on the dock • [class name] on the rocks • [class name] in the sunset • [class name] with a kite • [class name] on the sky • [class name] is on flight • [class name] is on flies	1. Landbird • landbird 2. Waterbird • waterbird

- **Modify** the cue by adding a keyword, e.g., “[class]'s photo” in [group], where the keyword represents the name of the group

- Obtaining the **average prompts embedding** for a class in all groups
- **Comparing broader** class embeddings for image classification

[4] Kim Y et al. Discovering and Mitigating Visual Biases through Keyword Explanation. CVPR, 2024.

Aspect 1 — Keyword

➢ **Application: Label diagnosis**

Keyword	Bee	Boar	Desk	Market
Samples				
Label	fly	pig	computer mouse	custard apple
Pred.	bee	wild boar	desktop computer	grocery store
Caption	a **bee** on a yellow flower.	wild **boar** in the forest.	the **desk** in the office.	fruit and vegetables at the **market**.

- B2T can diagnose common labeling errors, such as **mislabeling** and **label ambiguities**

[4] Kim Y et al. Discovering and Mitigating Visual Biases through Keyword Explanation. CVPR, 2024.

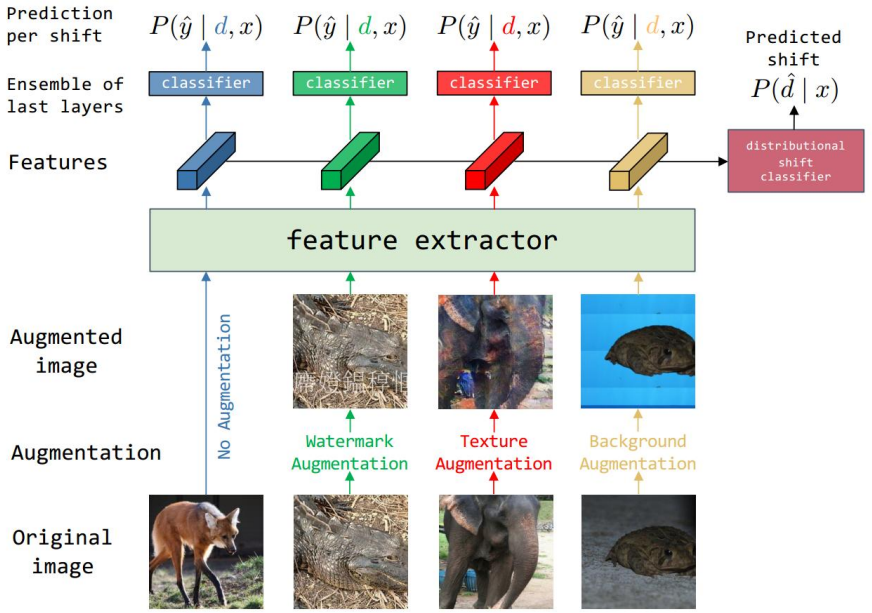
A Whac-A-Mole Dilemma: Shortcuts Come in Multiples Where Mitigating One Amplifies Others

†Zhiheng Li[2] *Ivan Evtimov[1] Albert Gordo[1] Caner Hazirbas[1] Tal Hassner[1]
Cristian Canton Ferrer[1] Chenliang Xu[2] *Mark Ibrahim[1]
[1]Meta AI [2]University of Rochester

Aspect 2 — Image-Level

➢ **Method**: LLE: Last Layer Ensemble

An ensemble of multiple classification layers (last layers) and a distributional shift classifier



Step1: Data Augmentation Strategies

- Predefine multiple augmentation types (e.g., **watermark augmentation**, **texture augmentation**, **background augmentation**) to simulate different shortcut biases

Step2: Train an Ensemble of Last Layers

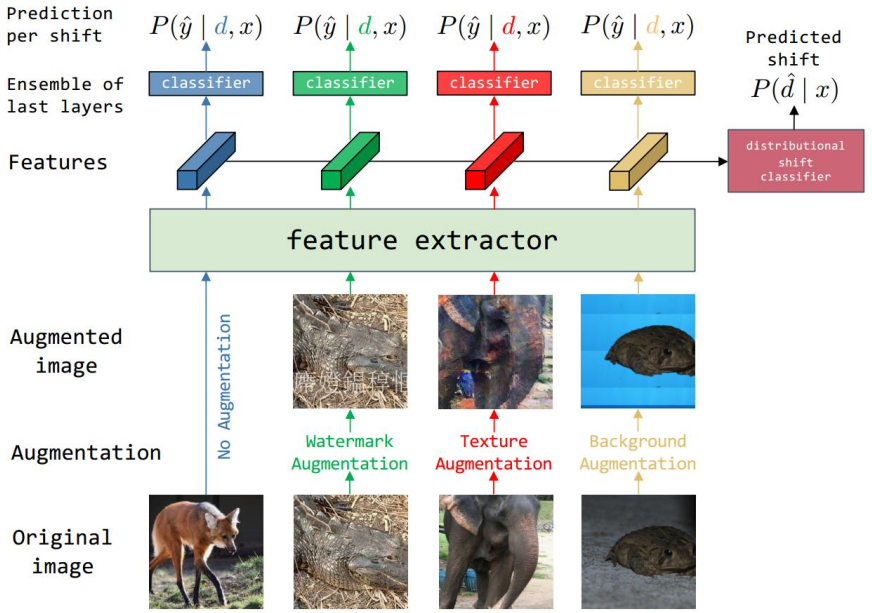
- Multiple classification layers are trained on a shared **feature extractor**, each focusing on specific augmentation type to mitigate particular bias.

[4] Li Z, Evtimov I, et al. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. CVPR, 2023.

Aspect 2 — Image-Level

➢ **Method**: LLE: Last Layer Ensemble

An ensemble of multiple classification layers (last layers) and a distributional shift classifier



Step3: Train a Distributional Shift Classifier

- To **predict** input's **augmentation type**
- Its **gradients are stopped** as for feature extractor

Step4: Dynamically Aggregate Predictions

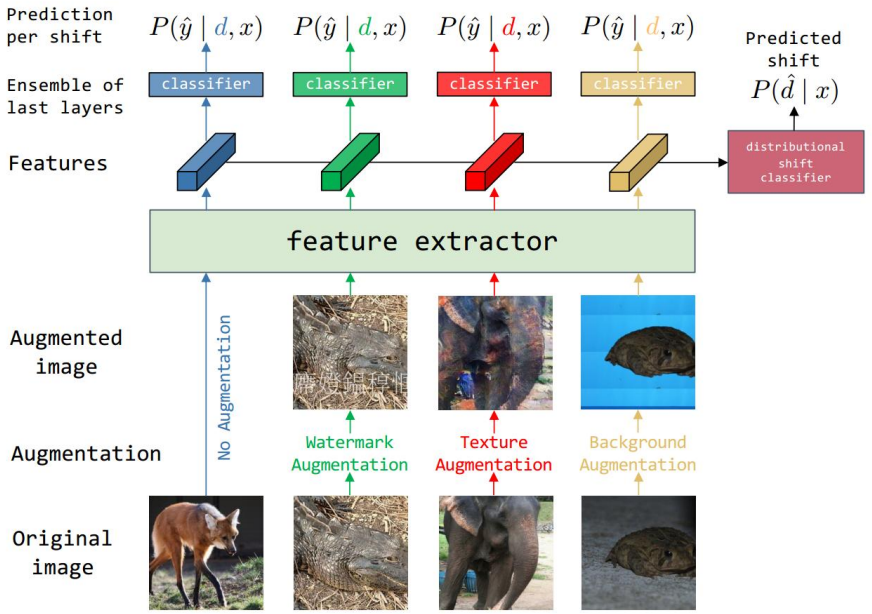
- During inference, the distributional shift classifier **dynamically weights** predictions **to reduce interference from irrelevant layers**.

[4] Li Z, Evtimov I, et al. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. CVPR, 2023.

Aspect 2 — Image-Level

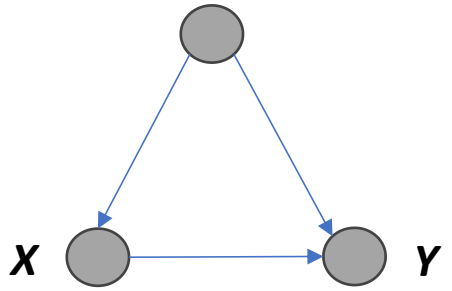
➢ **Interpretation** :

This is not a form of intervention, but more like dynamic weighting after image enhancement based on counterfactuals



Aspect	Conditional Probability (Dynamic Weighting)	Interventional Form
Formula	$P(\hat{y} \mid x) = \sum_d P(\hat{y} \mid d, x) P(d \mid x)$	$P(\hat{y} \mid x) = \sum_d P(\hat{y} \mid d, x) P(d)$
Weight Distribution Source	Dynamic distribution $P(d \mid x)$, dependent on input image x	Fixed distribution $P(d)$, representing global augmentation types
Dependency on Input Image	Weights dynamically depend on the input image x	Weights do not depend on the input image, based on global priors

{d,…}(confounder/shortcut)



[4] Li Z, Evtimov I, et al. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. CVPR, 2023.

Aspect 2 — Image-Level

➢ **Experiment** :

overcome **multiple shortcuts** or struggle in a **Whac-A-Mole behavior**

	I.D. Acc	shortcut reliance		
		BG Gap ↑	CoObj Gap ↑	BG+CoObj Gap ↑
ERM	97.6	-15.3	-11.2	-69.2
Mixup	98.3	-12.6	-9.3	-61.8
CutMix	96.6	-45.0 (×2.94)	-4.8	-86.5
Cutout	97.8	-15.8 (×1.03)	-10.4	-71.4
AugMix	98.2	-10.3	-12.1 (×1.08)	-70.2
SD	97.3	-15.0	-3.6	-36.1
CF+F Aug	96.8	-16.0 (×1.04)	**+0.4**	-19.4
LfF	97.2	-11.6	-18.4 (×1.64)	-63.2
JTT (E=1)	95.9	-8.1	-13.3 (×1.18)	-40.1
EIIL (E=1)	95.5	-4.2	-24.7 (×2.21)	-44.9
JTT (E=2)	94.6	-23.3 (×1.52)	-5.3	-52.1
EIIL (E=2)	95.5	-21.5 (×1.40)	-6.8	-49.6
DebiAN	98.0	-14.9	-10.5	-69.0
LLE (ours)	96.7	**-2.1**	-2.7	**-5.9**

On UrbanCars, methods without using shortcut labels

	IN-1k	shortcut reliance					
		Watermark		Texture		Background	
		IN-W ↑ Gap	Carton ↓ Gap	SIN ↑ Gap	IN-R ↑ Gap	IN-9 ↑ Gap	
arch: RG-32gf							
ERM	80.88	-14.15	+32	**-69.27**	-52.43	**-6.40**	
SEER (FT,IG-1B)	83.35	**-6.50**	**+18**	-73.04 (×1.05)	**-50.42**	-7.14 (×1.11)	
arch: ViT-B/32							
ERM	75.92	-8.71	+34	**-57.16**	-49.45	**-6.86**	
Uniform Soup (FT,WIT)	79.96	-7.90	+24	-59.67 (×1.04)	**-27.51**	-7.78 (×1.13)	
Greedy Soup (FT,WIT)	81.01	**-6.47**	**+16**	-59.61 (×1.04)	-30.01	-7.21 (×1.05)	
arch: ViT-B/16							
ERM	81.07	-6.69	+26	-62.60	-50.36	-5.36	
SWAG (LP,IG-3.6B)	81.89	-7.76 (×1.16)	+18	-67.33 (×1.08)	**-19.79**	-10.39 (×1.94)	
SWAG (FT,IG-3.6B)	85.29	-5.43	+24	-66.99 (×1.07)	-29.55	-4.44	
MoCov3 (LP)	76.65	-16.0 (×2.39)	+22	-63.36 (×1.01)	-56.86 (×1.12)	-7.80 (×1.45)	
MAE (FT)	83.72	-4.60	+24	-65.20 (×1.04)	-47.10	-4.45	
MAE+LLE (ours)	83.68	**-2.48**	**+6**	**-58.78**	-44.96	**-3.70**	
arch: ViT-L/16 or 14							
ERM	79.65	-6.14	+34	-61.43	-53.17	-6.50	
SWAG (LP,IG-3.6B)	85.13	-5.73	**+6**	-60.26	-10.17	-7.26 (×1.12)	
SWAG (FT,IG-3.6B)	88.07	-3.16	+20	-63.45 (×1.03)	-12.29	-2.92	
CLIP (zero-shot,WIT)	76.57	-4.47	+12	-61.27	**-6.26**	-3.68	
CLIP (zero-shot,LAION)	72.77	-4.94	+12	-56.85	-8.43	-4.54	
MAE (FT)	85.95	-4.36	+22	-62.48 (×1.02)	-36.46	-3.53	
MAE+LLE (ours)	85.84	**-1.74**	+12	**-56.32**	-34.64	**-2.77**	

On ImageNet, methods using self-supervised and foundation models

PERCEPTIONCLIP: VISUAL CLASSIFICATION BY INFERRING AND CONDITIONING ON CONTEXTS

Bang An[1*] **Sicheng Zhu**[1*] **Michael-Andrei Panaitescu-Liess**[1]
Chaithanya Kumar Mummadi[2] **Furong Huang**[1]

[1]University of Maryland, College Park [2]Bosch Center for Artificial Intelligence

Aspect 3 — Prompt-Level

➢ **Method**: PerceptionCLIP

Zero-shot inference

$$\texttt{CLIP}_1(y;x) \triangleq \langle \phi_I(x), \quad \phi_T(\alpha(y)) \rangle$$

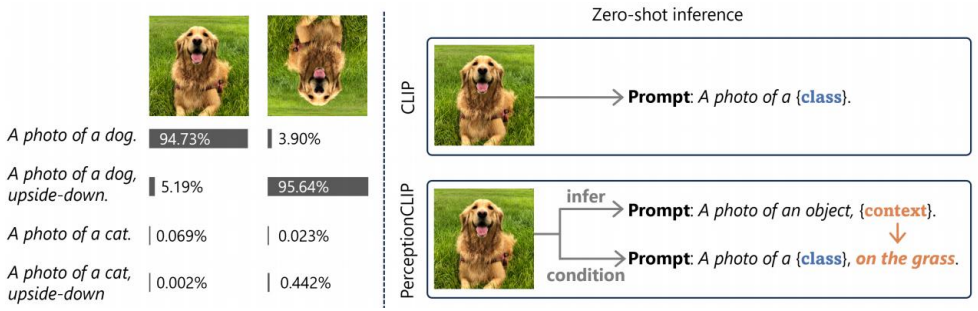
$$\texttt{CLIP}_{80}(y;x) \triangleq \left\langle \phi_I(x), \quad \frac{\frac{1}{80}\sum_{i=1}^{80}\phi_T(\alpha_i(y))}{\left\|\frac{1}{80}\sum_{i=1}^{80}\phi_T(\alpha_i(y))\right\|} \right\rangle$$

$$\texttt{CLIP}(y, z_1, \ldots, z_m; x) \triangleq \left\langle \phi_I(x), \quad \frac{\mathbb{E}\ \phi_T\big(\alpha(y) \oplus \alpha(z_1) \oplus \cdots \oplus \alpha(z_m)\big)}{\|\mathbb{E}\ \phi_T\big(\alpha(y) \oplus \alpha(z_1) \oplus \cdots \oplus \alpha(z_m)\big)\|} \right\rangle$$

Annotation function

$\alpha(y)$ = "a photo of a {class name of y}"

80 manually-designed templates $\{\alpha_i\}_{i=1}^{80}$



new description distribution:
$\alpha(y) \oplus \alpha(z1) \oplus \alpha(z2) \oplus \ldots$
e.g. "a photo of a dog, upright, bright"

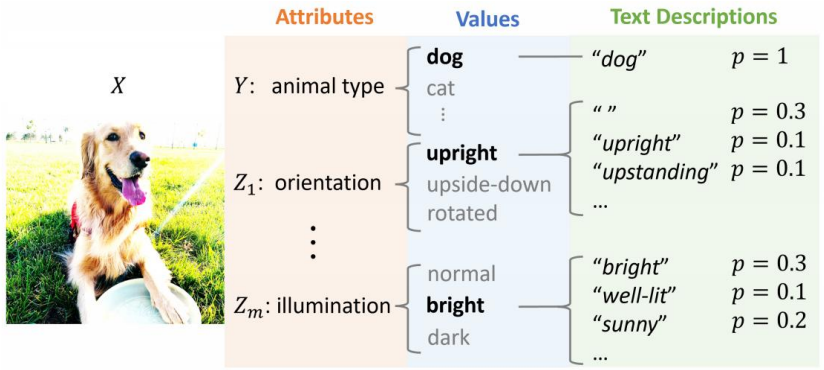
[5] An, Bang, et al. PerceptionCLIP: Visual Classification by Inferring and Conditioning on Contexts. ICLR, 2024.

Aspect 3 — Prompt-Level

➢ **Method**: PerceptionCLIP

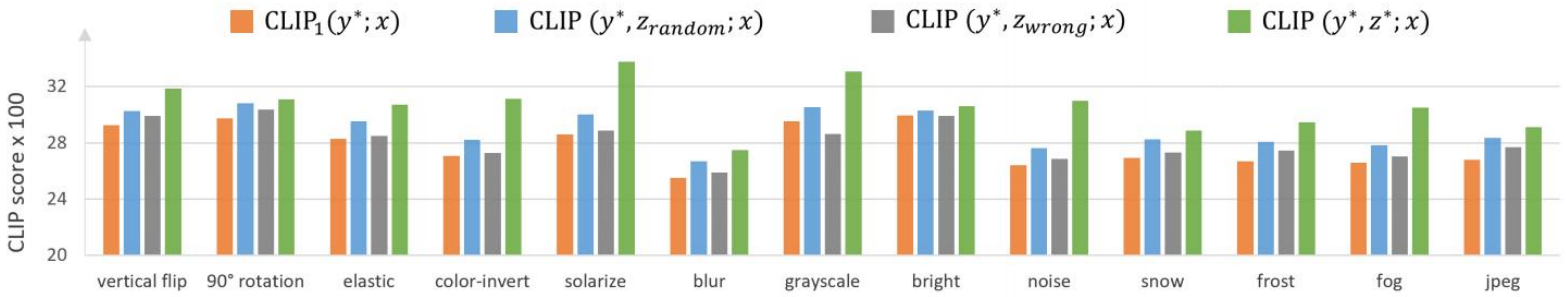
$$Y \rightarrow X \leftarrow \{Z_i\}_{i=1}^{m}$$

Step1: Inferring Contextual Attributes



- These attributes are viewed as **independent causal generative factors** that influence data generation.
- Prompt Construction and compute the conditional distribution $P(z_1, ..., z_m|x)$

$$\hat{p}(z_1, \ldots, z_m|x) \leftarrow \frac{\sum_y e^{\text{CLIP}(y, z_1, \ldots, z_m; x)/\tau}}{\sum_y \sum_{z_1, \ldots, z_m} e^{\text{CLIP}(y, z_1, \ldots, z_m; x)/\tau}} \text{ or } \frac{e^{\text{CLIP}(z_1, \ldots, z_m; x)/\tau}}{\sum_{z_1, \ldots, z_m} e^{\text{CLIP}(z_1, \ldots, z_m; x)/\tau}}$$



[5] An, Bang, et al. PerceptionCLIP: Visual Classification by Inferring and Conditioning on Contexts. ICLR, 2024.

Aspect 3 — prompt-level

➢ **Method**: PerceptionCLIP

Step2: Construct the final classification

- Contextual Conditional Distribution

$$\hat{p}(z_1,\ldots,z_m|x) \leftarrow \frac{\sum_y e^{\text{CLIP}(y,z_1,\ldots,z_m;x)/\tau}}{\sum_y \sum_{z_1,\ldots,z_m} e^{\text{CLIP}(y,z_1,\ldots,z_m;x)/\tau}} \text{ or } \frac{e^{\text{CLIP}(z_1,\ldots,z_m;x)/\tau}}{\sum_{z_1,\ldots,z_m} e^{\text{CLIP}(z_1,\ldots,z_m;x)/\tau}}$$

ClassAttr: "a photo of a {class name of y}, {description of z}."

PureAttr: "a photo of an object, {description of z}."

Better generalization

- Contextual Conditional Distribution

$$P(y \mid x, z_1, \ldots, z_m) = \frac{\exp(e_{\text{CLIP}}(y, z_1, \ldots, z_m; x)/\tau)}{\sum_{y'} \exp(e_{\text{CLIP}}(y', z_1, \ldots, z_m; x)/\tau)}$$

- Optimization Goals

$$\hat{y} = \text{argmax}_y \sum_{z_1,\ldots,z_m} P(y \mid x, z_1, \ldots, z_m) \cdot \hat{P}(z_1, \ldots, z_m \mid x)$$

[5] An, Bang, et al. PerceptionCLIP: Visual Classification by Inferring and Conditioning on Contexts. ICLR, 2024.

Aspect 3 — prompt-level

➢ **Experiment** :

Table 7: Average accuracy and worst group accuracy on the Waterbirds dataset.

	RN50			ViT-B/32			ViT-B/16			ViT-L/14		
	Avg ↑	**Worst ↑**	**Gap↓**	**Avg ↑**	**Worst ↑**	**Gap ↓**	**Avg ↑**	**Worst ↑**	**Gap ↓**	**Avg ↑**	**Worst ↑**	**Gap ↓**
without $\mathcal{Z}$	90.47	16.07	74.40	87.34	47.28	40.06	87.34	26.79	60.56	90.55	44.64	45.91
$\mathcal{Z}$={background}	88.78	16.07	72.71	89.80	66.07	23.73	82.98	16.07	66.91	86.44	44.94	41.51
$\mathcal{Z}$={background$^+$}	90.32	35.71	**54.61**	78.60	60.33	**18.28**	85.80	41.07	**44.73**	87.74	61.12	**26.62**

Table 8: Average accuracy and worst group accuracy on the CelebA dataset.

	RN50			ViT-B/32			ViT-B/16			ViT-L/14		
	Avg ↑	**Worst ↑**	**Gap↓**	**Avg ↑**	**Worst ↑**	**Gap ↓**	**Avg ↑**	**Worst ↑**	**Gap ↓**	**Avg ↑**	**Worst ↑**	**Gap ↓**
without $\mathcal{Z}$	81.05	73.87	7.19	80.73	75.82	4.91	75.16	62.01	13.16	86.98	77.36	9.61
$\mathcal{Z}$={gender}	85.10	80.44	4.65	79.89	76.70	**3.19**	75.27	65.13	10.14	80.30	74.31	5.99
$\mathcal{Z}$={gender, age}	87.71	84.98	**2.74**	82.82	78.06	4.76	75.81	65.52	10.29	82.26	79.06	3.21
$\mathcal{Z}$={gender, age, race}	85.55	82.51	3.05	82.02	75.94	6.09	77.17	69.18	**7.99**	83.04	80.84	**2.20**

[5] An, Bang, et al. PerceptionCLIP: Visual Classification by Inferring and Conditioning on Contexts. ICLR, 2024.

Aspect 3 — Prompt-Level

➢ **Interpretation** :

CLIP has the ability of Compositional Generalization

Disentangling object and attribute embeddings through contrastive learning ⟶ maximizes the mutual information

$$I(x_1, x_2; y_1, y_2) = \mathrm{D_{KL}}(p(x, y) \parallel p(x)p(y)) \xrightarrow{\text{Maximizing}}$$

- $x = (x_1, x_2)$: Image embeddings (objects and attributes).
- $y = (y_1, y_2)$: Text embeddings (objects and attributes).

- x_1 with y_1 (object embeddings).
- x_2 with y_2 (attribute embeddings).
- Enforces independence: $x_1 \perp x_2$, $y_1 \perp y_2$.

$$\begin{aligned}
I(x_1, x_2; y_1, y_2) &= \mathrm{D_{KL}}(p(x, y) \parallel p(x)p(y)) \\
&= \mathrm{D_{KL}}(p(x_1|x_2, y)p(x_2|y)p(y) \parallel p(x_1|x_2)p(x_2)p(y)) \\
&= \mathbb{E}_{x_1,x_2,y}[\log(p(x_1|x_2, y)/p(x_1|x_2))] + \mathbb{E}_{x_2,y}[\log(p(x_2|y)/p(x_2))] \\
&= \mathbb{E}_{x_2,y}[\mathrm{D_{KL}}(p(x_1|x_2, y) \parallel p(x_1|x_2))] + \mathbb{E}_y[\mathrm{D_{KL}}(p(x_2|y) \parallel p(x_2))]
\end{aligned}$$

[5] An, Bang, et al. PerceptionCLIP: Visual Classification by Inferring and Conditioning on Contexts. ICLR, 2024.

INTERPRETING CLIP'S IMAGE REPRESENTATION VIA TEXT-BASED DECOMPOSITION

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt
UC Berkeley
{yossi_gandelsman,aaefros,jsteinhardt}@berkeley.edu

Aspect 4 — Mechanism-Level

➢ **Method**: TEXTSPAN

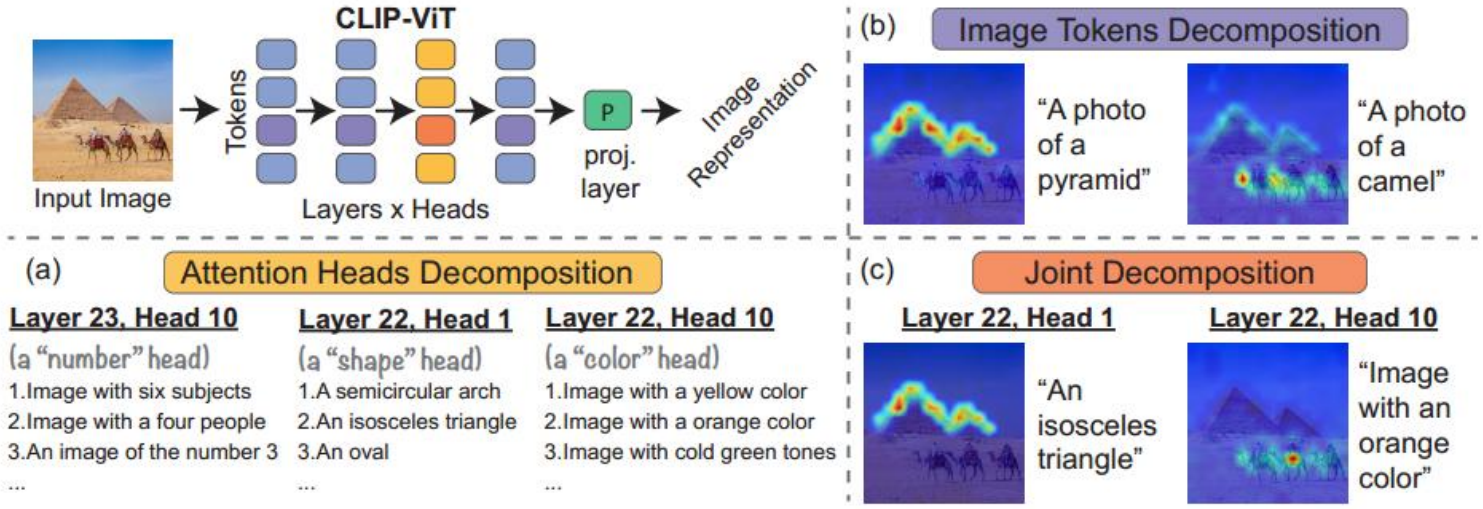


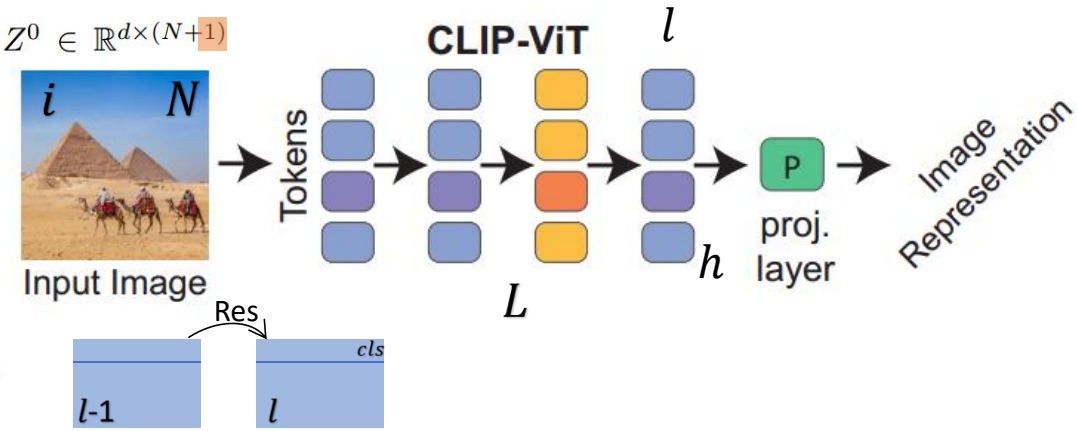
Figure 1: **CLIP-ViT image representation decomposition.** By decomposing CLIP's image representation as a sum across individual image patches, model layers, and attention heads, we can (a) characterize each head's role by automatically finding text-interpretable directions that span its output space, (b) highlight the image regions that contribute to the similarity score between image and text, and (c) present what regions contribute towards a found text direction at a specific head.

[6] Gandelsman Y, Efros A A. Interpreting CLIP's Image Representation via Text-Based Decomposition. ICLR, 2024.

Aspect 4 — Mechanism-Level

➢ **Method**: TEXTSPAN

$$M_{\text{image}}(I) = P\text{ViT}(I)$$



$$\hat{Z}^l = \text{MSA}^l(Z^{l-1}) + Z^{l-1}, \quad Z^l = \text{MLP}^l(\hat{Z}^l) + \hat{Z}^l.$$

$$M_{\text{image}}(I) = P\text{ViT}(I) = P\left[Z^0\right]_{cls} + \underbrace{\sum_{l=1}^{L} P\left[\text{MSA}^l(Z^{l-1})\right]_{cls}}_{\text{MSA terms}} + \underbrace{\sum_{l=1}^{L} P\left[\text{MLP}^l(\hat{Z}^l)\right]_{cls}}_{\text{MLP terms}}$$

direct effect

[6] Gandelsman Y, Efros A A. Interpreting CLIP's Image Representation via Text-Based Decomposition. ICLR, 2024.

Aspect 4 — Mechanism-Level

➢ **Method**: TEXTSPAN

$$M_{\text{image}}(I) = P\text{ViT}(I) = P\left[Z^0\right]_{cls} + \underbrace{\sum_{l=1}^{L} P\left[\text{MSA}^l(Z^{l-1})\right]_{cls}}_{\text{MSA terms}} + \underbrace{\sum_{l=1}^{L} P\left[\text{MLP}^l(\hat{Z}^l)\right]_{cls}}_{\text{MLP terms}}$$

direct effect

	Base accuracy	+ MLPs ablation
ViT-B-16	70.22	67.04
ViT-L-14	75.25	74.12
ViT-H-14	77.95	76.30

Table 1: **MLPs mean-ablation.** We simultaneously replace all the direct effects of the MLPs with their average taken across ImageNet's validation set. This results in only a small reduction in zero-shot classification performance.

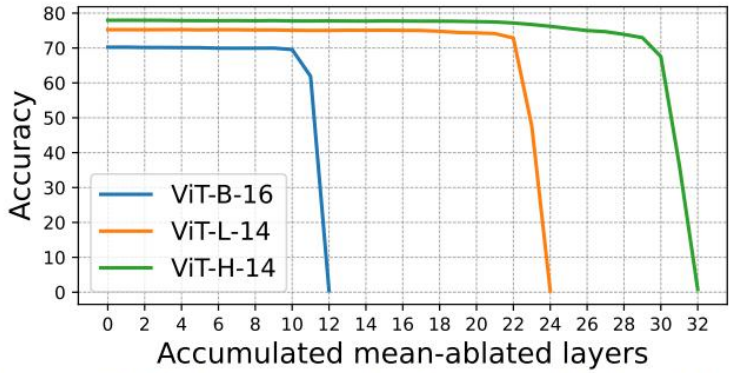


Figure 2: **MSAs accumulated mean-ablation.** We replace all the direct effects of the MSAs up to a given layer with their average taken across the ImageNet validation set. Only the replacement of the last few layers causes a large decrease in accuracy.

- **MLPs have a negligible direct effect**
- **Only the last MSAs have a significant direct effect**

[6] Gandelsman Y, Efros A A. Interpreting CLIP's Image Representation via Text-Based Decomposition. ICLR, 2024.

Aspect 3 — Mechanism-Level

➢ **Method**: TEXTSPAN

$$\sum_{l=1}^{L} P\left[\mathrm{MSA}^{l}(Z^{l-1})\right]_{cls} = \sum_{l=1}^{L}\sum_{h=1}^{H}\sum_{i=0}^{N} c_{i,l,h}, \quad c_{i,l,h} = P x_i^{l,h}$$

$$x_i^{l,h} = \alpha_i^{l,h} W_{VO}^{l,h} z_i^{l-1}$$

Attention weights from the class token to the i -th token

- **Decomposition into heads**

$$c_{\text{head}}^{l,h} = \sum_{i=0}^{N} c_{i,l,h}$$

- **Decomposition into tokens**

$$c_{\text{token}}^{i} = \sum_{l=1}^{L}\sum_{h=1}^{H} c_{i,l,h}$$

[6] Gandelsman Y, Efros A A. Interpreting CLIP's Image Representation via Text-Based Decomposition. ICLR, 2024.

Aspect 3 — Mechanism-Level

- **LN's perspective**

Decomposition incorporates: $LN(x) = \frac{\gamma}{\sqrt{\sigma^2+\epsilon}} \cdot x - \frac{\mu\gamma}{\sqrt{\sigma^2+\epsilon}} + \beta$

where multiplicative term $\frac{\gamma}{\sqrt{\sigma^2+\epsilon}}$ **absorbed into the projection matrix *P*,** while **additive terms are split across contributions**

Each input to MLPs and MSAs is layer-normalized before being processed:

$$\hat{Z}^l = \mathrm{MSA}(LN^l(Z^{l-1})) + Z^{l-1}, \quad Z^l = \mathrm{MLP}(LN^l(\hat{Z}^l)) + \hat{Z}^l$$

$$\left[\mathsf{MSA}^l(Z^{l-1})\right]_{cls} = \sum_{h=1}^{H}\sum_{i=0}^{N} x_i^{l,h}, \quad x_i^{l,h} = \alpha_i^{l,h}\mathsf{LN}^l(z_i^{l-1})W_{VO}^{l,h}$$

[6] Gandelsman Y, Efros A A. Interpreting CLIP's Image Representation via Text-Based Decomposition. ICLR, 2024.

Aspect 3 — Mechanism-Level

➢ **Method**: TEXTSPAN

- **Decomposition into heads** $c_{\text{head}}^{l,h} = \sum_{i=0}^{N} c_{i,l,h}$
- **Greedy algorithm for descriptive set mining**

$$V_{\text{explained}}(\mathcal{T}) = \frac{1}{K}\sum_{k=1}^{K} \| \text{Proj}_{\mathcal{T}}(c_k - c_{\text{avg}})\|_2^2, \text{ where } c_{\text{avg}} = \frac{1}{K}\sum_{k=1}^{K} c_k.$$

Algorithm 1: TEXTSPAN

Input: Head (l, h) contribution $c_{\text{head}}^{l,h}$ for K images stacked as rows in a matrix $C \in \mathbb{R}^{K\times d'}$, a pool of M text descriptions $\{t_i\}_{i=1}^{M}$, their corresponding CLIP text representations $R \in \mathbb{R}^{M\times d'}$ (projected to the head output space), and basis size m

Output: A set of text descriptions $\mathcal{T}$ and projected representations $C' \in \mathbb{R}^{K\times d'}$

Initialization: $C' \leftarrow \mathbf{0}_{K\times d'}$, $\mathcal{T} \leftarrow \phi$

for *i in* $[1, ..., m]$ **do**

- $D \leftarrow RC^T$
- $j^* \leftarrow \arg\max_{j=1}^{M} \text{Var}(D[j])$
- $\mathcal{T} \leftarrow \mathcal{T} \cup \{t_{j^*}\}$
- **for** *k in* $[1, ..., K]$ **do**
  - $C'[k] \leftarrow C'[k] + \frac{\langle C[k], R[j^*]\rangle}{||R[j^*]||^2} R[j^*]$
  - $C[k] \leftarrow C[k] - \frac{\langle C[k], R[j^*]\rangle}{||R[j^*]||^2} R[j^*]$
- **for** *k in* $[1, ..., M]$ **do**
  - $R[k] \leftarrow R[k] - \frac{\langle R[k], R[j^*]\rangle}{||R[j^*]||^2} R[j^*]$

principle component search

$\langle M_{\text{text}}(t), c_{\text{head}}^{l,h}\rangle$

[6] Gandelsman Y, Efros A A. Interpreting CLIP's Image Representation via Text-Based Decomposition. ICLR, 2024.

Aspect 3 — Experiment

➢ **Method**: TEXTSPAN $\langle M_{\text{text}}(t), c_{\text{head}}^{l,h} \rangle$

- **Decomposition into heads** $c_{\text{head}}^{l,h} = \sum_{i=0}^{N} c_{i,l,h}$

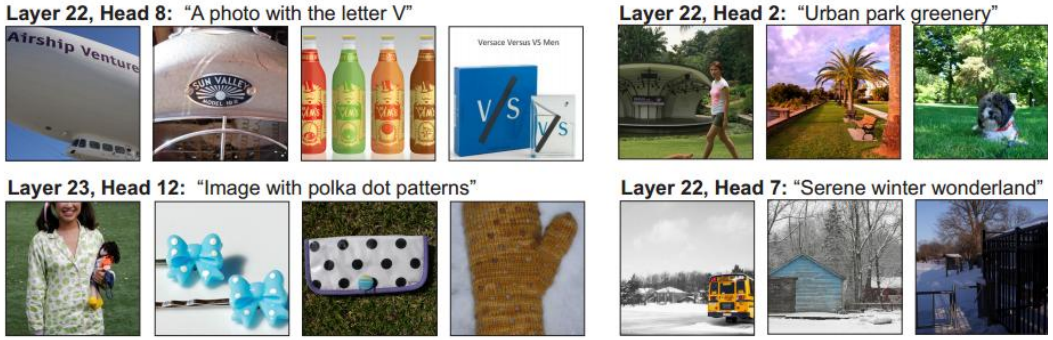


Figure 4: **Top-4 images for the top head description found by TEXTSPAN.**

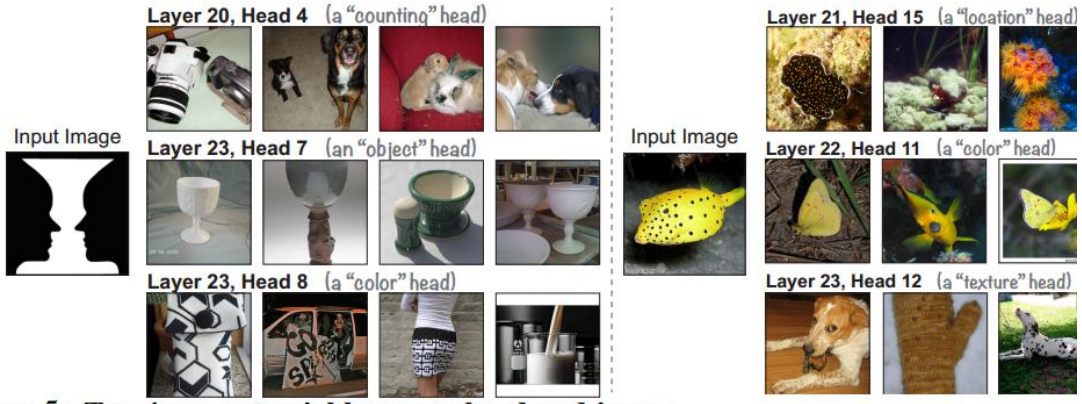


Figure 5: **Top-4 nearest neighbors per head and image.**

Layer 20, Head 0	**Layer 20, Head 1**
Picture taken in Hungary	Picture taken in Seychelles
Image taken in New England	Picture taken in Saudi Arabia
Futuristic technological concept	Muted urban tones
Playful siblings	Man-made pattern
Picture taken in the English countryside	an image of glasgow
Layer 20, Head 2	**Layer 20, Head 3**
Image of a police car	Intrica wood carvingte
Picture taken in Laos	Image snapped in Spain
Remote alpine chalet	Photo taken in Bora Bora, French Polynesia
A photograph of a small object	An image of a Preschool Teacher
Desert sandstorm	A breeze
Layer 20, Head 4	**Layer 20, Head 5**
Image with a pair of subjects	an image of samoa
Image with five subjects	Urban nostalgia
Image with a trio of friends	A photo with the letter K
A photo of an adult	Image snapped in the Colorado Rockies
Image with a seven people	Serendipitous discovery
Layer 20, Head 6	**Layer 20, Head 7**
Bustling city square	Energetic children
Peaceful village alleyway	Grumpy facial expression
ornate cathedral	Intricate ceramic patterns
Image taken in the Alaskan wilderness	Photo taken in Bangkok, Thailand
Modern airport terminal	Subdued moments
Layer 20, Head 8	**Layer 20, Head 9**
Photo taken in Rioja, Spain	Tranquil Asian temple
Photo taken in Borneo	Vibrant city nightlife
Vibrant urban energy	A photo with the letter R
Picture captured in the Icelandic glaciers	intricate mosaic artwork
serene oceanside scene	Photo taken in the Rub' al Khali (Empty Quarter)
Layer 20, Head 10	**Layer 20, Head 11**
A bowl	Photo taken in Beijing, China
A bottle	Photo with retro color filters
Nostalgic pathways	Image with holographic cyberpunk aesthetics
A laptop	Urban street fashion
Reflective ocean view	Photograph with the artistic style of tilt-shift
Layer 20, Head 12	**Layer 20, Head 13**
Photo with grainy, old film effect	Image taken from a distance
Detailed illustration	Photograph with the artistic style of split toning
Serene beach sunset	Photo taken in Beijing, China
An image of the number 10	A close-up shot
An image of the number 5	An image of a Novelist
Layer 20, Head 14	**Layer 20, Head 15**
Quirky street performer	Remote hilltop hut
Antique sculptural element	Photo taken in Barcelona, Spain
Celebratory atmosphere	Dynamic movement
Overwhelmed facial expression	Caricature of an influential leader
Serene winter wonderland	A picture of Samoa

[6] Gandelsman Y, Efros A A. Interpreting CLIP's Image Representation via Text-Based Decomposition. ICLR, 2024.

Aspect 3 — Mechanism-Level

➢ **Method**: TEXTSPAN $\langle c^i_{\text{token}}(I), M_{\text{text}}(t) \rangle$

- **Decomposition into tokens** $c^i_{\text{token}} = \sum_{l=1}^{L} \sum_{h=1}^{H} c_{i,l,h}$

$$M_{\text{image}}(I) = \sum_{i=0}^{N} c^i_{\text{token}}(I)$$

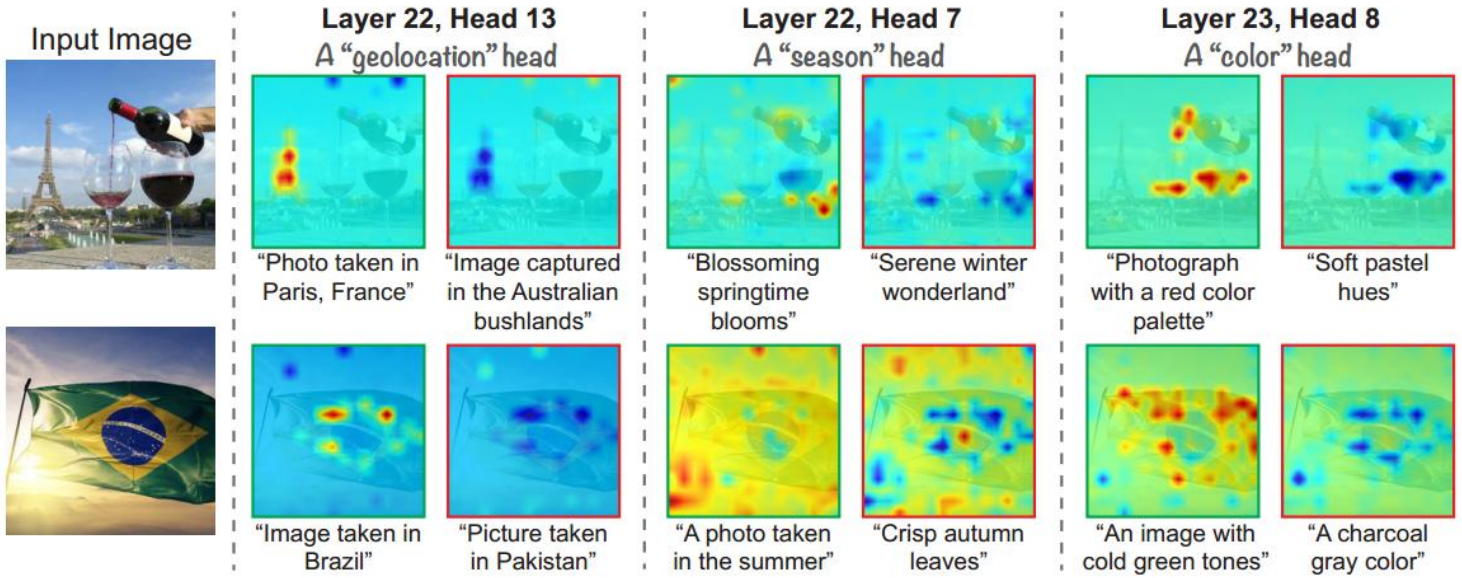


Figure 6: **Joint decomposition examples.** For each head (l, h), the left heatmap (green border) corresponds to the description that is most similar to $c^{l,h}_{\text{head}}$ among the TEXTSPAN output set. The right heatmap (red border) corresponds to the least similar text in this set (for $m = 60$). See Figure 9 for more results.

[6] Gandelsman Y, Efros A A. Interpreting CLIP's Image Representation via Text-Based Decomposition. ICLR, 2024.

Aspect 3 — Mechanism-Level

➢ **Method**: TEXTSPAN $\langle c^i_{\text{token}}(I), M_{\text{text}}(t) \rangle$

- **Decomposition into tokens** $c^i_{\text{token}} = \sum_{l=1}^{L} \sum_{h=1}^{H} c_{i,l,h}$

Figure 8: **Heatmaps produced by the image token decomposition**. We visualize (a) what areas in the image directly contribute to the similarity score between the image representation and a text representation and (b) what areas make an image representation more similar to one text representation rather than another.

[6] Gandelsman Y, Efros A A. Interpreting CLIP's Image Representation via Text-Based Decomposition. ICLR, 2024.

THANKS