Concept Regions Matter: Benchmarking CLIP with a New Cluster-Importance Approach

Aishwarya Agarwal Srikrishna Karanam Vineet Gandhi

{aishagar, skaranam}@adobe.com, vgandhi@iiit.ac.in

Adobe

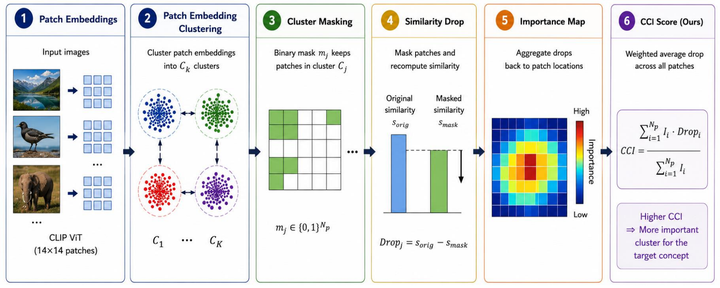
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

CVPR JUNE 3-7, 2026 DENVER COLORADO

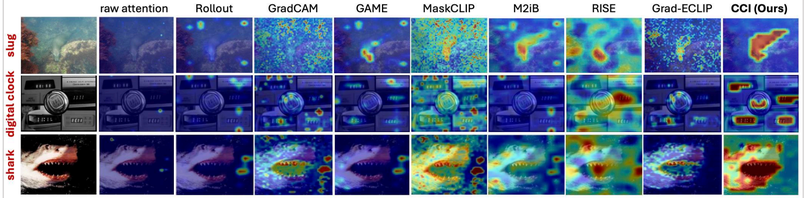
Motivation

- CLIP often relies on spurious background correlations (e.g. water → water ouzel)
- Accuracy alone is an unreliable proxy
  - correct predictions can still be background-driven
  - errors can rise from foreground-focused scenarios
- Existing dataset (CounterAnimals) uses only accuracy to define hard/easy sets
- We need a better interpretability method and a controlled benchmark to truly diagnose background reliance

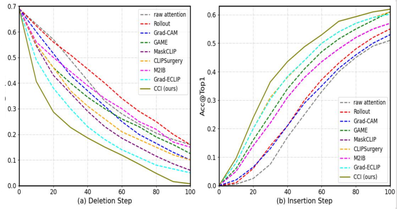
Concept Cluster Importance (CCI)



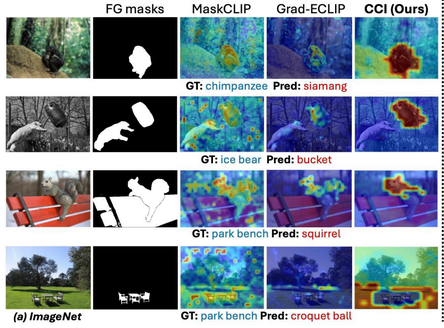
CCI Results

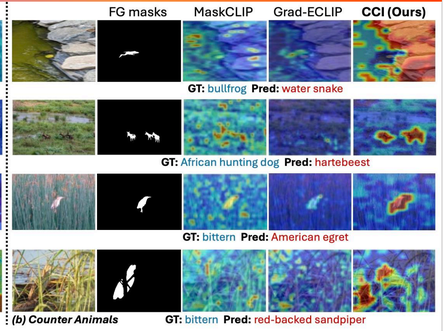


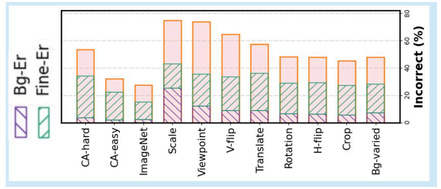
Method	Deletion ↓ Ground-truth @1	Deletion ↓ Ground-truth @5	Deletion ↓ Prediction @1	Deletion ↓ Prediction @5	Insertion ↑ Ground-truth @1	Insertion ↑ Ground-truth @5	Insertion ↑ Prediction @1	Insertion ↑ Prediction @5
raw attention	0.3831	0.6239	-	-	0.2492	0.4195	-	-
Rollout	0.4082	0.6556	-	-	0.2803	0.4665	-	-
Grad-CAM	0.3417	0.5628	0.3518	0.5817	0.2682	0.4454	0.2526	0.4206
GAME	0.3356	0.5734	0.3497	0.5938	0.3611	0.5636	0.3425	0.5384
MaskCLIP	0.2848	0.4885	0.2886	0.4957	0.3335	0.5351	0.3275	0.5267
CLIPSurgery	0.3115	0.5235	0.3217	0.5412	0.3832	0.6021	0.3727	0.5719
M2IB	0.3630	0.5953	0.3633	0.5951	0.3351	0.5411	0.3347	0.5410
Grad-ECLIP w/o λ_i	0.2535	0.4379	0.2634	0.4568	0.3715	0.5831	0.3528	0.5556
Grad-ECLIP	0.2464	0.4272	0.2543	0.4420	0.3838	0.5993	0.3672	0.5749
CCI (Ours)	**0.1809**	**0.3276**	**0.1789**	**0.3318**	**0.4175**	**0.6518**	**0.3893**	**0.6201**



CCI analysis of CLIP failures







- CCI reveals that many CLIP failures are driven by fine-grained object confusion or distraction objects in image rather than background reliance alone
- Background-driven errors constitute only a small fraction of overall prediction failures across existing datasets

COVAR: COntrolled VARiations Benchmark

We introduce controlled variations to disentangle background reliance from other factors driving CLIP failures

Evaluating CLIP variants on COVAR

Name	Method Patch	Res	Dataset	Bg-varied Acc	H-flip Acc	Translate Acc	Crop Acc	V-flip Acc	Rotation Acc	Viewpoint Acc	Scale Acc	Avg Acc
ViT-B	32	256	DataComp-1B	55.8	55.2	55.8	59.4	32.2	44.6	28.5	24.2	44.5
ViT-B	16	224	DataComp-1B	55.7	54.3	53.7	56.5	37.5	44.9	27.7	27.2	44.7
ViT-L	14	224	DataComp-1B	62.2	61.3	60.7	62.3	48.3	54.9	30.3	32.6	51.6
ViT-L	14	224	LAION-2B	59.7	58.9	60.0	60.2	43.2	53.9	32.2	30.1	49.8
ViT-H	14	224	LAION-2B	60.2	59.2	59.4	60.9	45.3	54.2	30.1	31.0	50.0
ViT-bigG	14	224	LAION-2B	61.6	62.1	62.2	63.8	45.7	56.5	34.2	33.5	52.5
ViT-L	14	224	DFN-2B	59.1	57.8	58.0	59.5	41.8	51.3	28.0	31.1	48.3
ViT-SO-SigLIP2	14	224	WebLI	63.3	62.7	62.4	63.6	53.0	58.6	34.1	36.9	54.3
ViT-B-SigLIP	16	256	WebLI	57.5	56.8	56.6	57.4	38.7	48.9	27.9	29.1	46.6
ViT-B-SigLIP2	16	384	WebLI	64.1	63.3	63.4	63.8	48.1	59.9	31.0	37.8	53.9
ViT-B-SigLIP2	32	256	WebLI	54.8	53.9	55.0	58.2	30.9	41.6	26.0	25.6	43.2
ViT-B-SigLIP2	16	512	WebLI	64.9	63.8	63.9	64.1	49.3	60.6	30.5	36.4	54.2
ViT-H-qgelu	14	378	DFN-5B	65.2	65.4	65.5	65.7	55.0	61.7	36.7	39.3	56.8
ViT-H-qgelu	14	224	DFN-5B	63.2	62.9	63.3	64.1	50.9	57.4	35.9	38.2	54.5
ViT-B	16	224	OpenAI	52.2	52.3	51.7	54.8	35.2	42.6	26.1	25.0	42.5
ViT-B	32	224	OpenAI	47.9	48.0	46.0	48.3	26.8	30.1	21.8	22.1	36.4
ViT-L	14	224	OpenAI	56.5	56.7	56.0	57.4	46.9	49.2	26.7	28.3	47.2
ViT-L	14	336	OpenAI	57.6	57.8	57.2	58.4	49.5	53.1	27.3	31.8	49.1

Name	Method Patch	Res	Dataset	Bg-varied BG-Er	Bg-varied Fine-Er	H-flip BG-Er	H-flip Fine-Er	Translate BG-Er	Translate Fine-Er	Crop BG-Er	Crop Fine-Er	V-flip BG-Er	V-flip Fine-Er	Rotation BG-Er	Rotation Fine-Er	Viewpoint BG-Er	Viewpoint Fine-Er	Scale BG-Er	Scale Fine-Er
ViT-B	32	256	DComp-1B	23.6	40.9	21.2	44.6	18.9	46.2	18.5	47.7	22.6	33.3	21.3	39.5	22.9	29.5	50.7	16.1
ViT-B	16	224	DComp-1B	14.2	49.3	11.7	52.7	11.3	53.4	11.9	52.2	13.8	42.6	12.4	50.4	15.7	33.5	30.5	28.4
ViT-L	14	224	DComp-1B	15.7	51.8	13.5	56.0	12.4	56.6	12.2	56.6	15.2	50.2	14.5	53.9	17.1	36.0	30.2	31.2
ViT-L	14	224	LAION-2B	16.9	49.8	14.0	55.0	13.8	53.2	12.4	55.4	15.4	44.3	16.2	51.4	17.1	33.5	35.5	27.8
ViT-H	14	224	LAION-2B	16.4	54.8	13.6	59.4	13.0	58.7	15.1	56.3	15.9	49.6	14.8	55.2	18.2	35.1	33.8	30.9
ViT-bigG	14	224	LAION-2B	17.8	51.2	16.0	54.3	16.4	53.6	17.2	52.3	15.4	48.3	17.4	54.0	19.6	33.0	32.7	29.7
ViT-L	14	224	DFN-2B	15.7	50.2	13.7	53.2	14.2	54.1	15.5	52.3	15.2	45.6	13.5	54.4	17.4	34.7	29.6	31.8
ViT-SO-SigLIP2	14	224	WebLI	25.3	47.8	22.7	52.8	21.2	53.1	20.0	53.8	23.0	51.7	25.4	51.4	27.0	32.7	42.3	29.3
ViT-B-SigLIP	16	256	WebLI	16.5	49.2	13.5	53.6	13.0	53.8	13.0	54.4	15.2	44.3	14.4	52.6	17.1	33.8	33.6	26.7
ViT-B-SigLIP2	16	384	WebLI	19.8	51.5	17.5	55.2	15.9	56.7	16.6	55.2	19.7	49.9	20.2	54.5	25.4	33.1	36.8	28.3
ViT-B-SigLIP2	32	256	WebLI	36.0	33.6	35.8	35.0	37.2	34.6	31.2	42.7	36.1	24.5	32.3	24.7	34.7	20.8	45.7	7.7
ViT-B-SigLIP2	16	512	WebLI	20.3	51.8	18.1	55.2	16.7	55.8	17.2	53.1	20.3	50.1	20.2	53.8	24.8	31.9	39.3	22.3
ViT-H-qgelu	14	378	DFN-5B	15.9	54.4	13.4	57.8	13.4	57.8	12.7	58.5	14.9	54.4	15.4	57.1	16.8	36.3	30.4	33.1
ViT-H-qgelu	14	224	DFN-5B	16.2	51.3	13.8	55.7	13.6	54.8	14.3	54.2	15.6	49.2	15.0	52.5	17.4	34.7	29.4	23.2
ViT-B	16	224	OpenAI	15.6	43.7	13.8	48.2	14.0	48.8	13.1	48.0	14.2	37.9	16.0	45.1	16.8	31.3	33.9	23.7
ViT-B	32	224	OpenAI	14.0	37.3	12.0	41.5	13.7	42.0	12.6	46.3	13.2	30.2	15.3	34.4	14.5	26.4	38.2	21.4
ViT-L	14	224	OpenAI	17.1	44.3	14.9	48.2	15.2	48.1	16.8	45.2	15.6	44.6	16.8	46.0	16.7	34.3	30.1	12.1
ViT-L	14	336	OpenAI	15.3	46.4	12.8	50.8	13.3	49.7	11.7	50.9	13.4	48.1	15.3	49.2	14.4	35.4	28.4	30.4

- Larger models such as **ViT-H-qgelu (DFN-5B)** and **ViT-bigG** achieve the highest overall accuracies on COVAR, with ViT-H-qgelu(378px) reaching **56.8% average accuracy** across all subsets
- **Scale** and **viewpoint** variations are the most challenging perturbations across all CLIP variants, causing substantially larger drops than flips, translation, or crop transformations
- Models trained on curated datasets such as **DataComp-1B** exhibit lower background reliance compared to large-scale noisy web data models like LAION-based variants
- Smaller patch-size models (e.g., **ViT-B/16**) consistently show lower BG-Er than **ViT-B/32**, suggesting finer-grained patch representations help reduce shortcut reliance

Key Takeaway

- Accuracy alone is insufficient for diagnosing background bias in CLIP
- Fine-grained confusion and robustness shifts contribute more to failures than background reliance alone
- CCI combined with COVAR enables deeper diagnosis of CLIP robustness behaviour

Paper