Seeing Beyond the Scene: Analyzing and Mitigating Background Bias in Action Recognition

Ellie Zhou

Issue Statement

1. **Background Bias** is when models rely on the background, rather than the human motion to make predictions. This bias arises because datasets contain consistent correlations between actions and backgrounds (e.g., skiing is always on snow), leading models to leverage backgrounds as shortcuts for the action.

Model's Prediction
Playing baseball: 0.78
Playing violin : 0.12
. . .

Model's Prediction
Shopping: 0.81
Dancing: 0.09
. . .

2. **Real-Life Implications**

Healthcare	AI relies on hospital-specific watermarks or medical equipment rather than disease symptoms in X-rays, leading to unreliable diagnostics.
Public Safety	Self-driving cars learn to use crosswalks as shortcut for pedestrian, causing it to fail to recognize human when crosswalk is missing.
Law Enforcement	Surveillance AI uses specific neighborhood backgrounds as shortcut for criminal activity, leading to individuals being targeted based on surroundings rather than behavior.

Research on mitigating background bias is the **foundation** to creating trustworthy AI models that focus on the task, not shortcuts.

Related Work

HAT [1]	• Evaluates background bias in a wide range of classification models • **Gap**: does not propose solutions to mitigate bias
Why Can't I Dance in the Mall? [2]	• Mitigates bias in CNN classification models • **Gap**: does not extend evaluation or mitigation to more powerful and modern models like Video-LLMs
Removing the Background by Adding the Background [3]	• Mitigates bias in CNN classification models • **Gap:** does not extend to other models, does not quantify background reliance

Research Question and Objective

Question	How do I quantify background bias? How much do modern model paradigms rely on background scene shortcuts? How can I design solutions to effectively mitigate these biases?
Objective	• Systematically quantify background bias across classification models, contrastive image-text models, and VLLMs • Develop solutions to reduce background reliance without compromising overall accuracy

Video Datasets and Metrics

Playing trumpet in **baseball field**

HAT Action-Swap [1]: synthetic video dataset of humans doing action A pasted on background of different action B.
- ***Human Accuracy (HuAcc)***: # of times predict human action / total
- ***Background Error (BgErr)***: # of times predict background action / total

We want HuAcc to be **high** and BgErr to be **low**.

Mimetics: video dataset containing mimed actions, so it lacks background context (eg: surfing in house)
- ***Accuracy***: # of times predict human action / total

Kinetics: standard benchmark video dataset containing humans performing various actions (eg: cooking sausages)
- ***Accuracy***: # of times predict human action / total

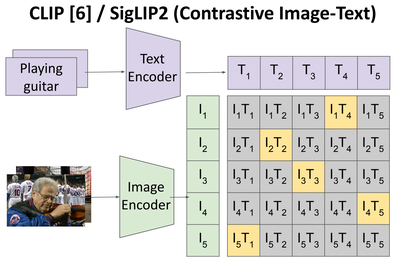
Analysis of Background Bias in Video Models

Slow-Only [4] (Classification)

RNN Layers

Output Logits
Hitting Baseball - 0.55
Playing Violin - 0.30
…
Playing guitar - 0.01
Bowling - 0.01

CLIP [6] / SigLIP2 (Contrastive Image-Text)



VLLM

Prompt: What is the action being performed?

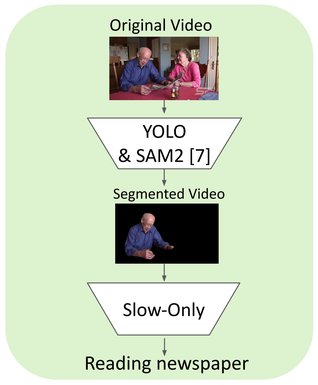
VLLM

Playing violin

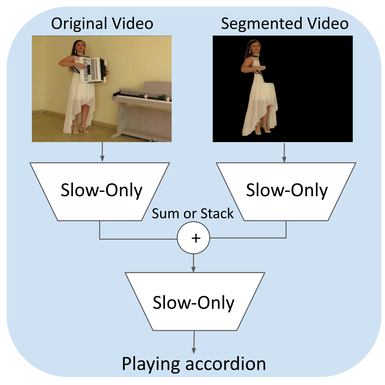
	Model	HAT (w/ bg-swap) HuAcc ↑	HAT (w/ bg-swap) BgErr ↓	Mimetics (w/o bg-swap) Accuracy ↑
	All classes			
	Slow-Only	**11.71**	26.24	**6.31**
	CLIP ViT-B/32	4.32	**15.07**	5.75
	SigLIP2	4.06	20.01	4.63
	As 5-choice MCQ			
Classification	Slow-Only	35.81	55.41	57.64
Contrastive Image-Text	CLIP ViT-B/32	29.25	53.66	46.84
Contrastive Image-Text	SigLIP2	25.46	58.91	48.95
VLLM	InternVL3-8B	40.29	48.84	62.83
VLLM	InternVL3-78B	**45.73**	**48.39**	**66.61**

All models show background bias (BgErr > HuAcc), with Video-LLMs showing the least amount of background bias.

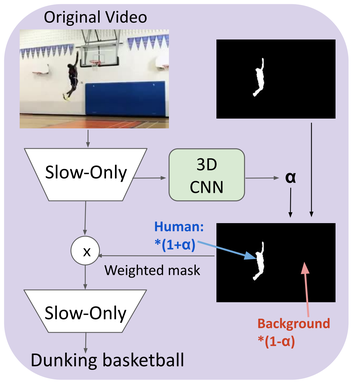
Four Architectural Innovations for Mitigating Bias in Classification Models



1 - Segmented



2, 3 - Dual Branch (Sum or Stack)



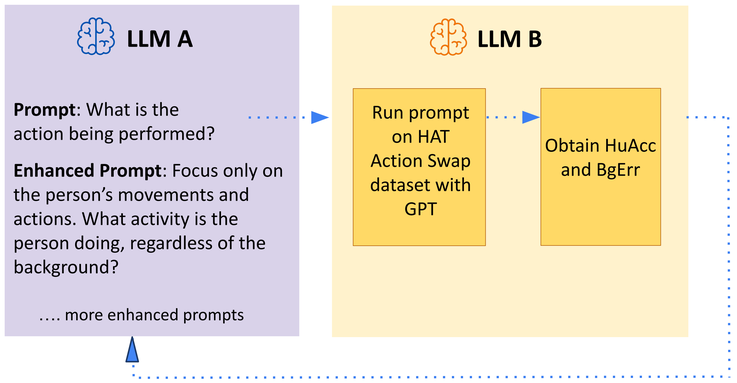
4 - Weighted-Focus

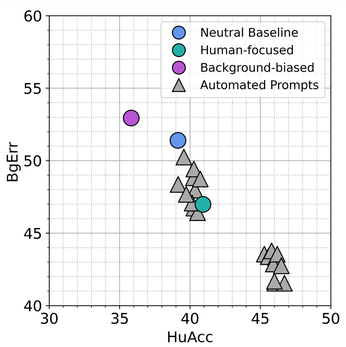
Model	HAT (background-swapped) HuAcc ↑	HAT (background-swapped) BgErr ↓	Mimetics Accuracy ↑	Kinetics Accuracy ↑
Slow-Only (baseline)	9.62	23.42	6.87	49.93
Segmented	23.34 (+13.72)	2.09 (-21.33)	9.54 (+2.67)	23.46 (-26.47)
Dual-Branch Sum	12.76 (+3.14)	20.36 (-3.06)	7.85 (+0.98)	52.15 (+2.22)
Dual-Branch Stack	12.80 (+3.18)	19.80 (-3.62)	8.28 (+1.41)	51.51 (+1.58)
Weighted-Focus	12.80 (+3.18)	19.64 (-3.78)	7.85 (+0.98)	52.03 (+2.10)

Feeding only human leads to no background bias, but results in huge performance drop on the standard benchmark Kinetics (see Segmented row).

However, if human segmentation is fed along with original video, there is both reduced bias and increase in standard benchmark performance (Kinetics).

Novel Automated and Iterative Prompt Design for Video Large Language Models (VLLMs)

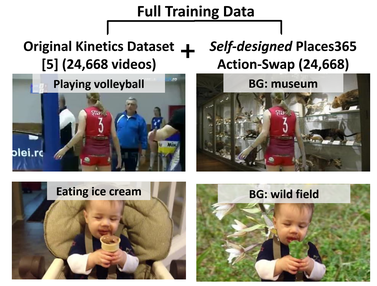




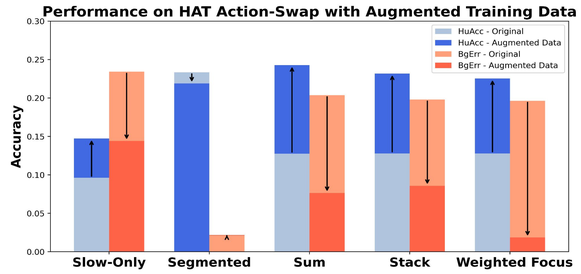
Prompt design can steer Video-LLM to have more or less background bias.

- **Worst Prompt**: Please just look at the background and not the person. Based on the background scene, what is the action being performed?
- **Best Prompt**: Ignore where the video takes place. What action is the person doing, based only on their movements?

Self-Designed Data Augmentation



Since background bias stems from scene-action correlations, I designed a targeted augmentation to break the associations and encourage the model to focus less on background.



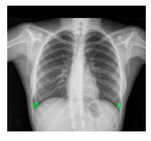
Augmentation increases HuAcc across most models and decreases BgErr, effectively mitigating background bias.

Achievements

1. I **quantified background bias** across classification models, contrastive image-text models, and Video-LLMs, finding all exhibit significant bias.
2. My classification model mitigations **solve the "fairness-accuracy" tradeoff** by reducing reliance on background cues while simultaneously improving accuracy on standard benchmarks.
3. For Video-LLMs, my **automated prompt-engineering solution** effectively reduces background bias without the need for expensive re-training.
4. Beyond **architectural innovations**, adding **self-designed** augmented training data reduces bias even more by encouraging the model to generalize across diverse synthetic environments.

Impact

As AI is becoming increasingly integrated in the world, it is imperative to make sure that their decision-making processes base their decisions on relevant causal actions rather than unrelated spurious correlations. This research provides a critical framework for enhancing AI robustness and equity, ensuring it remains reliable in high-stakes real-world deployments.



Future Work

- Incorporate the mitigation solutions into real-world applications such as remote patient monitoring, surveillance systems, or autonomous robotics
- Investigate presence of other biases such as gender, race, age, geographic, resource setting, etc. in the same model paradigms
- Experiment with bias mitigation in contrastive image-text models

Selected References

1. Jihoon Chung, Yu Wu, and Olga Russakovsky. Enabling detailed action recognition evaluation through video dataset augmentation. Advances in Neural Information Processing Systems, 35:39020–39033, 2022.
2. Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. Advances in Neural Information Processing Systems, 32, 2019.
3. Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, and Xing Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11804–11813, 2021.
4. Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6202–6211, 2019.
5. Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308, 2017.
6. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR, 2021.
7. Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024.