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

Ellie Zhou
Westmont High School
ellie.m.zhou@gmail.com

Jihoon Chung
Princeton University
jc5933@princeton.edu

Olga Russakovsky
Princeton University
olgarus@princeton.edu

Abstract

Human action recognition models often rely on background cues, rather than human movement and pose to make predictions, a behavior known as background bias. In this paper, we present a systematic analysis of background bias across classification models, Vision-Language models, and Large Language Models and find that all exhibit a strong tendency to default to background reasoning, though LLMs show relatively reduced reliance on background compared to the other two. Next, we propose background bias mitigation strategies for classification models and show that incorporating additional segmented human input effectively decreases background bias. Finally, we explore both manual and automated prompt tuning for large language models, demonstrating that prompt design can steer predictions towards human-focused reasoning.

1 Introduction

Human action recognition aims to identify what the human is doing in the video, but models often rely on the background, rather than human to make predictions, a phenomenon known as background bias [4, 5]. E.g., given a video of a person playing violin in a baseball field, the model may predict "playing baseball" rather than "playing violin" because of the background. This bias arises because datasets often contain consistent correlations between actions and backgrounds (e.g., skiing always occurs on snow), leading models to leverage backgrounds as shortcuts for the action [9].

Despite the growing popularity of Vision-Language models like CLIP [14], SigLIP2 [16], and multi-modal large language models [2, 12, 23], background bias in these model paradigms has not been extensively studied. Prior research [4, 5, 11, 17] has mainly focused on classification models. Some works have examined background bias in CLIP and video LLMs. [7, 18, 21] primarily focus on object classification rather than action recognition, employ simplistic removal-based strategies, and lack a systematic comparison of background bias across model paradigms. While MASH-VLM [1] does a great job analyzing scene bias, their work mostly tests on scene-only videos, making it hard to quantify how much models rely on human versus background context. Our work fills the gap by performing a comprehensive analysis of background bias across model paradigms while also developing architectural and prompting-based mitigation strategies.

The main contributions are:
(1) We analyze background bias across model paradigms including CLIP, SigLIP2, and LLMs. We find that all models display background bias, while LLMs rely less on background cues.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: What Makes a Good Video: Next Practices in Video Generation and Evaluation.

Table 1: Results of models on background bias benchmarks (left) and CLIP class-level analysis on HAT Action Swap (right). Left: all models show background bias, though LLMs rely less on background cues. Right: high bias arises when backgrounds are distinctive and consistently paired with actions.

Model	SHAcc ↑	SBErr ↓	Mimetics ↑
All classes			
Slow-Only [6]	**11.71**	26.24	**6.31**
CLIP ViT-B/32 [14]	4.32	**15.07**	5.75
SigLIP2 [16]	4.06	20.01	4.63
As 5-choice MCQ			
Slow-Only [6]	35.81	55.41	57.64
CLIP ViT-B/32 [14]	29.25	53.66	46.84
SigLIP2 [16]	25.46	58.91	48.95
InternVL3-8B [23]	40.29	48.84	62.83
InternVL3-78B [23]	**45.73**	**48.39**	**66.61**

High Bias (highest SBErr)		
Background Class	SHAcc	SBErr
presenting weather forecast	3.64	89.09
decorating the christmas tree	4.62	75.38
cutting pineapple	0.00	70.37
ice skating	0.00	69.39
cleaning toilet	0.00	68.09
Low Bias (lowest SBErr)		
Background Class	SHAcc	SBErr
dancing ballet	10.87	0.00
playing clarinet	10.91	0.00
washing feet	10.91	0.00
waxing chest	14.81	0.00
laughing	15.52	0.00

(2) We propose strategies to mitigate background bias in classification models, finding that incorporating segmented human inputs reduces background bias.
(3) We demonstrate that prompt engineering can effectively steer LLMs toward human-focused reasoning, with automated prompt tuning emerging as a particularly promising approach.

2 Dataset and Metrics

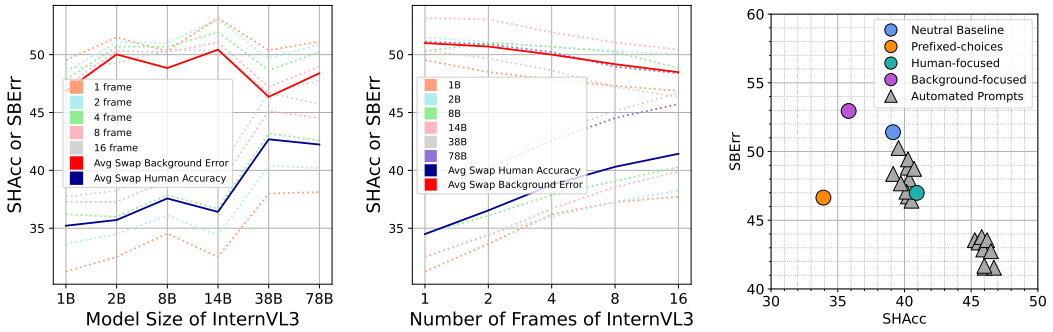
To measure background bias in a model, we use HAT Action Swap, introduced in [5], which contains videos where the human from class A is placed on a mismatched background of class B (e.g., eating ice cream on basketball court). We report Swap Human Accuracy (SHAcc) and Swap Background Error (SBErr). SHAcc is the fraction of videos where the model correctly predicted the human class (A), and SBErr is the fraction incorrectly predicted as the background class (B). A high SHAcc indicates more reliance on human features, while a high SBErr shows reliance on background. We also evaluate on Mimetics [19], which contains mimed actions without matching scene context, and Kinetics [3], a human action video dataset. We report accuracy for Mimetics and Kinetics.

3 Analysis of Background Bias

3.1 Vision-Language Models

We analyzed background bias in CLIP ViT-B/32 [14] and SigLIP2 [16] models, both contrastive image-text models known to perform well on a wide range of visual understanding tasks without task-specific training. We tested on the three "Random" Action-Swap mixes from HAT [5], which are generated from the Kinetics-400 dataset, by feeding CLIP and SigLIP2 the center video frame. The 400 action labels from Kinetics were used as text prompts. The action prediction was determined based on which text embedding had the highest similarity score with the image embedding.

As shown in Table 1 (left), both CLIP and SigLIP2 have a higher swap background error than swap human accuracy, indicating that they are biased towards predicting the background. We further broke down the results by background class label. Table 1 (right) shows the background category and their SHAcc and SBErr. For example, among all the videos which have "presenting weather forecast" as the background, only 3.64% predicted the correct human label, while 89.09% predicted "presenting weather forecast". In general, high background bias tends to occur when the background is visually distinctive and consistently paired with the action (e.g., christmas tree in decorating christmas tree, toilet in cleaning toilet). In such cases, the model tends to over-rely on background cues, predicting the background-associated action even when the person's movements do not match. On the other hand, low background bias tends to occur when the background is not distinctive for the action and could appear in many different contexts. For example, "playing clarinet" could take place in a concert hall, a living room, or outdoors, so there is no distinctive background cue that the model can rely on.



(a) (b) (c)

Figure 1: **(a)** Effect of model size on InternVL3. As model capacity increases, SHAcc improves, but SBErr persists. **(b)** Effect of number of frames on InternVL3. SHAcc increases, while SBErr decreases, showing temporal information helps. **(c)** Performance of GPT-4o-mini prompts on HAT Action Swap. Automated prompt tuning achieves higher accuracy and better reduces background bias.

Table 2: Result for Classification Model Mitigation Solutions. Change from Slow-Only baseline is shown in parenthesis.

Model	Kinetics-50 Acc ↑	HAT SHAcc ↑	HAT SBErr ↓	Mimetics Acc ↑
Slow-Only	49.93	9.62	23.42	6.87
Segmented	$23.46_{(-26.47)}$	$23.34_{(+13.72)}$	$2.09_{(-21.33)}$	$9.54_{(+2.67)}$
Dual-Branch Sum	$52.15_{(+2.22)}$	$12.76_{(+3.14)}$	$20.36_{(-3.06)}$	$7.85_{(+0.98)}$
Dual-Branch Stack	$51.51_{(+1.58)}$	$12.80_{(+3.18)}$	$19.80_{(-3.62)}$	$8.28_{(+1.41)}$
Weighted-Focus	$52.03_{(+2.10)}$	$12.80_{(+3.18)}$	$19.64_{(-3.78)}$	$7.85_{(+0.98)}$

3.2 Large Language Models

We conducted experiments on InternVL3 [23], a recent multimodal language model capable of inputting videos to investigate whether LLMs also exhibit background bias in action recognition. The task was framed as a multiple-choice question. We first prompted the LLM with the prompt: "What is the action being performed?" and then provided it with five answer choices. The choices included the human action label, background action label, and three randomly selected action classes from the remaining Kinetics-400 classes. To study performance trends, we varied (1) the number of frames (sampled evenly throughout the video) that we feed to the model and (2) the model size within the InternVL3 family.

As shown in Figure 1a, as the model size increases, the SHAcc improves. However, the Swap Background Error in HAT Action-Swap persists. This suggests that larger model capacity alone is insufficient for robust human action understanding. Figure 1b shows that as we give more frames, the SHAcc increases, while the SBErr decreases. This trend shows that temporal information helps the model focus more on the human motion than the background context.

Table 1 (left) tabulates the results of different models on background bias benchmarks. Overall, LLMs display less background bias than both classification model (Slow-Only) and Vision-Language models (CLIP and SigLIP2).

4 Mitigating Background Bias

In this section, we explore solutions to mitigate background bias in classification models and LLMs. For this section, instead of using Kinetics-400, we constructed Kinetics-50 using the 50 classes from Mimetics, and similarly, we construct mini HAT Action Swap, which uses the same 50 classes.

4.1 Classification Models

As a baseline model, we used the Slow-Only (R50 backbone) model [6], which shows strong background bias as it can be seen from the first row of Table 2. To mitigate this background bias, we propose four strategies that either restrict the model's access to background information or rebalance the contribution of human and background features. Details are available in the appendix.

Segmented Input We remove the background and only put human segmented-video [10, 15] as input. **Dual-Branch: Sum and Stack** Original video and segmented videos are fed to different Slow-Only, where they are fused in the middle via (1) element-wise addition or (2) channel concatenation. **Weighted Focus** We allow the model to adaptively control human vs. background weighting.

Figure 2: Examples showing original HAT Action Swap frame, binary mask, and final human segmentation. Human and Background classes of the images are: (Row 1) reading book on catching or throwing baseball, (Row 2) playing trumpet on hitting baseball, (Row 3) playing saxophone on juggling balls, (Row 4) opening bottle on cleaning windows.

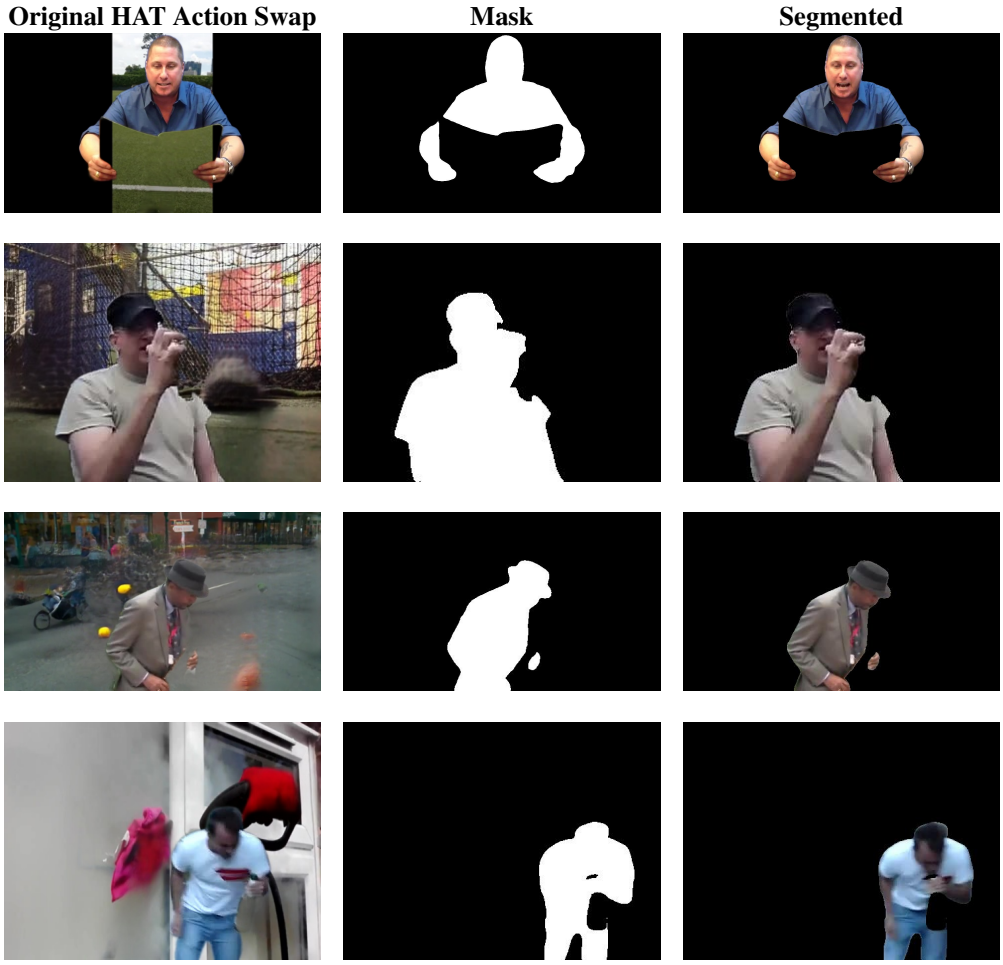


Figure 2 shows visual examples of action swap images, binary mask, and human segmentation. As shown in Table 2 (second row), feeding just the human alone leads to significant improvements in background bias, seen through the improvements in HAT SHAcc, HAT SBErr, and Mimetics. However, it suffers a large drop in Kinetics-50 performance, due to the lack of useful contextual cues, hinting that the background is still an important cue for action understanding. Dual-Branch Sum/Stack and Weighted Focus improve accuracy on Kinetics-50 while also lowering SBErr, increasing SHAcc, and increasing Mimetics accuracy. This shows that incorporating an additional segmented input can effectively improve performance and reduce background bias. These findings highlight an important

tradeoff: removing background reduces bias, but hurts accuracy when on datasets where action and background are strongly correlated, since useful context is lost. In contrast, retaining background preserves accuracy on such datasets but increases bias. Dual-Branch Sum/Stack and Weighted-Focus achieve an effective balance, improving Kinetics-50 accuracy while reducing background bias.

Additionally, we tried augmenting the training data where the background is replaced with an unrelated scene. Augmented data significantly improves performance on HAT Action-Swap, but with some sacrifice to the original Kinetics performance. Details of this is in the appendix.

4.2 Large Language Models

In this section, we explore prompt tuning as a strategy to mitigate background bias in LLMs. We used GPT-4o Mini [13], since it has strong general understanding of language, which would make it well-suited for evaluating different prompts. Evaluation is done on 75% of HAT Action Swap, and remaining 25% is used for automated prompt tuning.

Hand-Crafted Prompt Tuning We first tested hand-crafted prompts with varying levels of guidance. The prompts were as follows:
Neutral baseline - "*What is the action being performed?*" followed by five action labels.
Prefixed-choices - Same as neutral, but each choice is prefixed by "*a video of a human...* "
Human-focused - Instruct model to consider only the human while ignoring background.
Background-focused - Instruct model to consider only background while ignoring human.

As shown in Figure 1c, prompting can steer background bias. Specifically, we see that Human-focused prompts improves background bias upon the neutral prompt, while Background-focused worsens the bias.

Automated Prompt Tuning We showed that manual prompt engineering can reduce background bias to some extent. This raised the question: was this the limit of LLM performance, or simply a limitation of our prompts? To investigate, we turned to automated prompt engineering [8, 20], which systematically improves prompts through a feedback-driven loop using LLM. Figure 1c summarizes the performance of the four manually crafted prompts and the 20 automated prompts. While the human-focused manual prompt modestly improved SHAcc and SBErr relative to baseline, automated tuning consistently achieved larger gains, further increasing SHAcc and reducing SBErr. These results highlight prompt tuning as a more effective background bias mitigation strategy than manual prompt design.

5 Conclusion

In this study, we analyze background bias in action recognition across newer model paradigms - Vision-Language Models and LLMs - and find that both exhibit background bias, but with LLM showing the least amount of background bias. We then explore solutions for mitigating background bias in action recognition classification models and find that models which integrate both the original and segmented inputs improve performance on standard datasets like Kinetics and reduce background bias on counterfactual benchmarks like HAT Action-Swap. Turning to large language models, we show that they are sensitive to prompt wording and that background bias can be reduced through carefully designed manual prompts or automated prompt tuning.

References

[1] Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, and Jinwoo Choi. Mash-vlm: Mitigating action-scene hallucination in video-llms through disentangled spatial-temporal representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13744–13753, 2025.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen

Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Takumi Fukuzawa, Kensho Hara, Hirokatsu Kataoka, and Toru Tamaki. Can masking background and object reduce static bias for zero-shot action recognition? In *International Conference on Multimedia Modeling*, pages 366–379. Springer, 2025.

[8] Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Bianca Lamm, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. Are vision language models texture or shape biased and can we steer them? *arXiv preprint arXiv:2403.09193*, 2024.

[9] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[10] Glenn Jocher. Ultralytics yolov5, 2020. URL https://github.com/ultralytics/yolov5.

[11] Haoxin Li, Yuan Liu, Hanwang Zhang, and Boyang Li. Mitigating and evaluating static bias of action representations in the background and the foreground. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19911–19923, October 2023.

[12] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2024. URL https://arxiv.org/abs/2311.10122.

[13] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis,

Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

[14] 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.

[15] 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.

[16] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[17] 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.

[18] Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. A sober look at the robustness of clips to spurious features. *Advances in Neural Information Processing Systems*, 37:122484–122523, 2024.

[19] Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards understanding human actions out of context. *International Journal of Computer Vision*, 129(5):1675–1690, 2021.

[20] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*, 2023.

[21] Wenqian Ye, Guangtao Zheng, Yunsheng Ma, Xu Cao, Bolin Lai, James M Rehg, and Aidong Zhang. Mm-spubench: Towards better understanding of spurious biases in multimodal llms. *arXiv preprint arXiv:2406.17126*, 2024.

[22] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[23] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Technical Appendices and Supplementary Material

A.1 Construction of Kinetics-50 and mini HAT Action Swap

Kinetics-50 To construct the Kinetics 50 dataset, we began with the Kinetics-400 dataset [3] and selected the 50 action classes (same 50 as Mimetics [19]). For the training and validation data, we split the original Kinetics-400 training set (filtered to include only those 50 classes) into 80% for training and 20% for validation. For the testing data, we used the original Kinetics-400 validation set, again filtered to include only the 50 selected classes. The final dataset consists of 24,668 training videos, 6167 validation videos, and 2485 test videos.

Mini HAT Action Swap We use images from the HAT dataset introduced by [5]. They provide pre-generated segmented human videos and inpainted background videos for the Kinetics classes. We used the videos from same 50 classes as above to generate Action-Swap images, where a human figure is combined with a mismatched background drawn from a different action class. This counterfactual setup allowed us to evaluate whether models rely more on human appearance or scene context. The total size of our mini HAT Action-Swap set was 2366 videos.

A.2 Classification Model Details

Segmented Input As the most straightforward solution, we remove the background entirely by applying segmentation before feeding the video into Slow-Only [6]. This ensures the model sees only the human and cannot rely on background cues. We first used YOLOv5 [10] to extract object bounding boxes and their confidence scores. From these, the highest-confidence bounding box labeled with the "person" class was selected to be input for SAM2 [15]. We then used SAM2 to propagate the segmentation across all frames to extract consistent human segmentations throughout the video. Non-human regions were set to 0.

Dual-Branch While segmented input reduces reliance on background cues, it also removes potentially useful context. For example, if the model sees a person swimming, but no water, it may struggle to distinguish between swimming and other similar movements. Therefore, we created a dual-branch architecture that allows the model to learn dynamically from both human and background. The model consists of two parallel streams: one receiving original video and the other receiving segmented video (using same method as previous). Both inputs independently go through the initial layers of Slow-Only (Stem, Stage 1, and Stage 2). After Stage 2, the two feature maps are fused using one of two strategies: (1) Sum: element-wise channel addition, or (2) Stack: concatenation along the channel dimension. The fused representation goes through the remaining layers of the Slow-Only backbone (Stage 3, Stage 4, and Head). In the Stack method, Stage 3 is modified to accept the doubled channel dimension resulting from concatenation. The intuition is that in the early layers, each branch learns low-level features from its input, and after fused, it learns a joint representation integrating both human-focused and contextual cues.

Weighted Focus This approach focuses on letting the model adaptively control weighting between human and background. We introduce an auxiliary 3D CNN network that processes the early feature maps of the Slow-Only model. It learns a scalar parameter α, which controls relative human-background weighting. To apply this weighting, we use the binary segmentation mask M (where 1 = human, 0 = background) and compute a weighted mask as follows:

$$M_{\text{weighted}} = (1 + \alpha) \cdot M + (1 - \alpha) \cdot (1 - M) \tag{1}$$

To ensure stability, α is constrained to [-1, 1] using sigmoid, meaning the human region can be scaled up to $2\times$ or down to $0\times$, with the background scaled in the opposite direction. The weighted mask is then multiplied with the feature maps, and the modified features continue through the remaining layers of Slow-Only.

A.3 Classification Model Training Details

All classification models are based on the Slow-only (R50 backbone) variant of the SlowFast model [6]. We sampled 8 evenly spaced frames from each video and applied the same preprocessing transformations as proposed in the original SlowFast paper, including resizing and center-cropping to 224×224.

Model	Kinetics-50 ↑	Kinetics-50 (+aug) ↑	SHAcc ↑	SHAcc (+aug) ↑	SBErr ↓	SBErr (+aug) ↓	Mimetics ↑	Mimetics (+aug) ↑
Slow-Only [6]	49.94	46.56	9.62	14.73	23.42	14.41	6.87	10.24
Segmented	23.46	23.26	23.34	21.89	2.09	2.17	9.54	9.82
Dual-Branch Sum	52.16	39.72	12.76	24.27	20.36	7.65	7.85	9.12
Dual-Branch Stack	51.51	42.98	12.80	23.18	19.80	8.57	8.27	10.94
Weighted Focus	52.03	23.86	12.76	22.54	19.64	1.85	7.85	10.10

Table 3: Results of Classification Models with and without training on Places365 Augmented data. Adding augmented data drops performance on Kinetics-50, but improves performance on HAT Action Swap and Mimetics.

Models were all trained from scratch on Kinetics-50. Models were trained using the Adam optimizer with a starting learning rate of 0.001. We used a ReduceLROnPlateau scheduler with a patience of 40 epochs and a threshold of 1e-2 to adjust the learning rate dynamically based on validation loss. The batch size was set to 20 for all training runs. All models were trained for 300 epochs.

A.4 Augmented Data Experiment

We explored whether augmenting the training data could further improve performance of our classification models. In addition to the Kinetics-50 videos used to train all previous models, we created our own "action-swapped" videos - taking inspiration from the HAT Action-Swap set [5].

To construct these, we used YOLOv5 [10] and SAM2 [15] to extract human figures from each video in the Kinetics-50 training set. We then pasted each onto a background randomly selected from Places365 [22] - a large-scale dataset with 365 diverse scene categories. This allowed us to simulate context-mismatched scenarios where the action and background don't naturally align, challenging the model to rely less on background cues and focus more on the human motion. Each video kept the same background throughout, and we resized the background image to match the resolution of the original video. This data augmentation doubled the size of the training data, so the total number of training videos was 49,336 (= 24,668 $\times$ 2).

Adding augmented training data led to a trade-off in performance. As shown in Table 3, accuracy on Kinetics-50 dropped for all models when we trained on augmented data. However, performance improved on HAT Action Swap and Mimetics across all models. This suggests that the augmented data helped the model become more robust to background bias and context-mismatched scenarios - which are emphasized in HAT [5] and Mimetics [19]. On the other hand, since Kinetics typically has consistent background-action alignment, the original background cues may have been more helpful for classification, which explains the performance drop.

A.5 Manual and Automated Prompts for LLM

Manual Prompts

- What is the action being performed? *(same prompt for both neutral baseline and prefixed-choices)*
- Focus only on the person and their motion. Ignore the background, scene, or surroundings. Based on the person's posture, appearance, and movement, what is the action being performed?
- Please just look at the background and not the person. Based on the background scene, what is the action being performed?

Automated Prompts and Performance

Our automated engineering approach is inspired by the method introduced in [8, 20]. We simulate a back and forth conversation with GPT 4.1 [13], where it acts as a prompt engineer, rather than us. Below is how we simulated the conversation.

(1) We first instruct GPT to design a prompt to improve accuracy and reduce background bias. The instructions are the same as in [8], but adapted to suit our context - by replacing references to texture and shape with background and human cues.
(2) GPT responds with a proposed prompt (starting with PROMPT:).
(3) We test that prompt on our dataset using GPT 4o-mini and report back the human accuracy and

background error.
(4) GPT uses that feedback to refine its prompt in the next round.

Below are the 20 generated automated prompts and performance. Best performing prompt is bolded.

- Focus only on the person's movements and actions. What activity is the person doing, regardless of the background?
  SHAcc: 40.24%, SBErr: 48.83%
- Ignore the background. Based only on the person's movements, what action are they performing?
  SHAcc: 45.27%, SBErr: 43.55%
- Describe only the main action the person is doing, without considering the background or location.
  SHAcc: 40.26%, SBErr: 49.43%
- Based solely on the person's body movements, what action are they performing in this video? Ignore the background.
  SHAcc: 45.56%, SBErr: 43.38%
- Ignore the setting. What is the person doing, based only on their actions and movements?
  SHAcc: 46.07%, SBErr: 43.32%
- Disregard the background. Identify the action the person is performing by observing their movements only.
  SHAcc: 45.91%, SBErr: 42.87%
- Only consider the person's actions and body movements. What activity are they doing, without using any clues from the background?
  SHAcc: 39.15%, SBErr: 48.37%
- Focus only on the person's motion and behavior. What action are they performing, ignoring all background details?
  SHAcc: 46.02%, SBErr: 41.55%
- Watch the person's movements and actions only. What are they doing, without using any information from the background?
  SHAcc: 40.26%, SBErr: 46.74%
- Based only on the person's physical actions, what activity are they performing? Do not use any background information.
  SHAcc: 39.55%, SBErr: 50.26%
- Ignore everything except the person's movements. What action are they performing?
  SHAcc: 45.79%, SBErr: 43.78%
- Looking only at the person's actions, what are they doing in this video? Ignore the surroundings.
  SHAcc: 40.35%, SBErr: 47.97%
- Ignore the environment. What is the person doing, based only on their actions in the video?
  SHAcc: 46.19%, SBErr: 43.55%
- Focus only on the person's movements in the video. What action are they performing, without considering the background?
  SHAcc: 40.53%, SBErr: 46.42%
- **Ignore where the video takes place. What action is the person doing, based only on their movements?**
  SHAcc: 46.70%, SBErr: 41.55%
- Disregard the location and background. What is the person doing, based only on their actions?
  SHAcc: 46.48%, SBErr: 42.75%
- Without using any clues from the background or location, what action is the person performing in this video?
  SHAcc: 40.15%, SBErr: 47.08%

- Ignore the background and setting. What action is the person performing, based only on their movements?
  SHAcc: 45.99%, SBErr: 41.69%
- What is the person doing in this video, based only on their actions and not the background?
  SHAcc: 39.73%, SBErr: 47.68%
- Describe the action the person is performing, using only their movements and ignoring the background.
  SHAcc: 40.72%, SBErr: 48.74%

A.6 Limitations

While we have shown background bias in both Vision-Language models and video-LLMs, we have only tested on a handful of models and our results may not give the fuller picture of how other models might behave. Similarly, our methods of mitigating background bias were only tested on selected representable models, but we were not able to check if the conclusions hold true for different classification models. Automatic prompt tuning, specifically using LLM as an optimizer, is an actively researched area but still in its experimental stages. For example, while we have seen that some of the prompts were yielding improved results, the improvements were not incremental, where the following prompt is not always better than the previous prompt. While we have mainly used HAT as our background bias benchmark, as the original authors stated, the dataset is synthetically generated and might not follow real-world trends.