Same Action, Different Style: Can Vision-Language Models Tell How an Action Is Performed?

Kolja Hribar
Department of Computer Science
University of Pittsburgh
Pittsburgh, USA
koh16@pitt.edu

Adriana Kovashka
Department of Computer Science
University of Pittsburgh
Pittsburgh, USA
kovashka@cs.pitt.edu

***Abstract*—Vision-language models (VLMs) can tell *what* is happening in a video: the action, the objects, and the setting. But the same everyday action can be done in many different styles, and a lot of what sets one apart from another is *how* it is done, not *what* is done. We ask whether current VLMs can tell how an everyday action is performed, not just what is happening. To capture a wide range of styles, we take egocentric videos from the Ego4D dataset and pose 311 claims about how five actions are performed across six countries, then evaluate three VLMs on verifying these claims from the videos. We find that the models capture *what* is in the scene far better than *how* an action is performed, reading tool and object cues reliably but performing worse on the finer motion and grip cues that actually distinguish one style from another. They also tend to say yes to a claim that sounds right without checking the video, so their most common error is confirming a style that the clip never shows. Together, these results show that current models still miss the fine-motion detail that separates one style from another: the information a robot would need to imitate how a person performs a task.**

I. Introduction

Vision-language models (VLMs) can already describe much of *what* happens in a video: they can name the action, the objects, and the setting. But everyday manual actions are not performed in a single standard way. The same task is carried out in many different styles, and a lot of what sets one apart from another is *how* it is done, not *what* is done. Whether today's VLMs actually pick up on this has barely been tested. One place it matters is robots that take instructions in language: something like 'pour slowly' or 'scrub in circles' specifies style, and a model that cannot read the *how* cannot follow or copy it.

Most work on video and action understanding is about *what* action is happening. The goal is to put the right label on the action, and a model is judged by how often it gets that label right. Significantly less is known about whether a model can tell *how* the action was done: the technique, motion, rhythm, and grip that set one style apart from another. That is exactly where the stylistic differences show up, but it is easy to miss them when an action gets reduced to a single label. A limited number of works [1, 2] study variations in actions through adverbs, but adverbs are only a subset of the possible variations. Further, these works precede the strongest recent VLMs, so the latter's ability to detect variations was not previously studied.

To get a wide range of styles to test on, we use egocentric videos from the Ego4D dataset [3], which capture the same everyday actions done by many different people, in many different settings. We test three current VLMs (Qwen2-VL [4], VideoChat-R1.5 [5], and VideoLLaMA3 [6]) on whether they can tell how an action is performed. We pose 311 claims about how five actions (cutting, painting, cleaning, washing, pouring) are done across six countries (India, Italy, Japan, Saudi Arabia, UK, USA), and ask each model to verify these claims against human-labeled egocentric videos. Each claim is about either *what* is in the scene (the tools and objects) or *how* the action is done (its motion, rhythm, and grip). We find a clear gap: the models read the *what* reliably but do much worse on the *how*, and they tend to accept a claim that sounds right even when it does not match the video. In short, current VLMs recognize the action but miss the manner, which is exactly the part a robot would need in order to copy how it is done.

II. Related Work

Most work on video has gone toward naming the action that takes place. Action recognition has evolved from two-stream designs [7] and inflated 3D convolutions [8] to multi-pathway networks [9], and to transformer-based [10] and self-supervised approaches [11].

Closest to our question is work on adverbs, which describe how an action is done, such as cutting slowly or stirring gently. Moltisanti et al. [1] learn action changes from verb-adverb relationships, and Doughty and Snoek [2] recognize adverbs across different actions. Adverbs, though, capture only some of the ways an action can vary, and both came before the strong current VLMs, so how well today's models read manner was left open.

A related line of work looks at how the same action varies across people and settings. Plizzari et al. [12] build ARGO1M from Ego4D and show that recognition models generalize poorly when an action is performed in an unseen scenario and location, though their goal is still the action label across domains, not the manner within a clip.

Other work goes a step further and judges the execution itself. Rai and Kovashka [13] find that video-LLMs struggle to give useful feedback on how a person performs, using rock climbing as a case study, and make up for scarce feedback

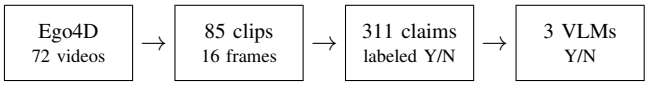


Fig. 1. Pipeline. Action intervals from 72 Ego4D videos are trimmed into 85 clips of sixteen frames each, paired with 311 hand-labeled style claims, and verified by three VLMs.

data with freely available video and text. Our work builds on these but asks a more direct question, whether current VLMs can read the manner of an everyday action at all, and where that reading breaks down.

III. Approach

Our goal is to measure whether current VLMs can tell *how* an action is performed, so we need clips of the same actions done in many styles, a set of concrete style claims to check, and a way to score each model against a human label. We build this in four stages: sampling clips from Ego4D, writing yes/no claims and verifying them by hand, asking each model to judge every claim from the frames, and scoring its answers against our labels. Fig. 1 summarizes the pipeline.

A. Data

We draw clips from Ego4D [3], a large collection of egocentric video in which many different people perform the same everyday tasks in their own settings, which gives us the range of styles we want to test. From the dataset's narrated action annotations, we select clips covering five manual actions (cutting, painting, cleaning, washing, and pouring) across six countries (India, Italy, Japan, Saudi Arabia, the UK, and the USA). Each selected action interval is trimmed from its source video into a short clip, and from every clip we sample sixteen evenly spaced frames, which are the input the models see. Because a video can contain more than one of our five actions, the 72 source videos produced 85 clips, which form the basis for the 311 claims we evaluate.

B. Style Claims

For each clip we pose a set of short claims about the action, phrased as yes/no statements about a single attribute (for example, "circular scrubbing motion" or "power grip on the tool"). Every claim is of one of two kinds: *what* is in the scene, covering the tools and objects, or *how* the action is done, covering its motion, rhythm, and grip. We label each claim against its clip as present (Y) or absent (N). One author labeled all 311 claims, and another independently labeled the 83 claims from two of the five actions (cutting and cleaning) as a reliability check. The two agree on 80.7% of these claims ($\kappa = 0.61$), but far more on *what* claims (92.7%, $\kappa = 0.85$) than on *how* claims (69.0%, $\kappa = 0.40$), so the motion and grip attributes are harder to pin down for people as well (though still easier for them than for any of the models). This gives us a ground-truth benchmark of 311 verified claims.

C. Verification

We evaluate three current VLMs: Qwen2-VL [4], VideoChat-R1.5 [5], and VideoLLaMA3 [6].[1] Each model sees the sixteen frames of a clip and one claim, and answers with a single character, Y if the claim holds and N if it does not. We ask for this yes/no decision rather than a free-form description we would have to score afterwards, which keeps every judgment tied to a specific attribute and directly comparable to the human label. The frames are identical across models, only the input format differs: one video block for Qwen2-VL and VideoChat-R1.5 and sixteen image blocks for VideoLLaMA3. Sixteen is a fixed budget for every clip, enough to capture a short action while keeping the input small and equal across models. We cap generation at eight tokens and read only the first character, which every model produced cleanly (all 933 answers were a valid Y or N).

D. Scoring

For every model we compare its answer on each claim against the human label and report accuracy as the fraction of claims on which the two agree. Because all of the claims are tagged with their action, country, and kind (*what* or *how*), we can break accuracy down along each of these to see not just how often a model is right, but where it succeeds and where it fails.

IV. Results

Table I reports each model's accuracy against the human labels. On plain accuracy the three land close together, between 69.8% and 72.3%. However, that number is misleading: the benchmark holds far more present claims than absent ones (204 to 107), so a model that simply answers "yes" every time already scores 65.6% without watching anything, and the three barely clear that floor. Balanced accuracy fixes this by scoring present and absent claims separately and averaging them, which pulls the always "yes" baseline down to 50%. Measured this way, all three models sit well above chance (63.4–68.3%), and Qwen2-VL pulls slightly ahead, though not significantly (exact McNemar, $p > 0.4$ for all three pairs). Two patterns explain the rest, and they run through the remainder of this section: the models read *what* far better than *how*, and confirm claims far more reliably than they reject them.

TABLE I
Overall and balanced accuracy of each model against the human labels across all 311 claims. Balanced accuracy is the mean of accuracy on present and absent claims.

Model	Accuracy (%)	Balanced acc. (%)
Qwen2-VL	72.3	68.3
VideoChat-R1.5	70.7	64.1
VideoLLaMA3	69.8	63.4
Majority-class baseline	65.6	50.0

[1]Checkpoints: Qwen/Qwen2-VL-7B-Instruct, OpenGVLab/VideoChat-R1_5-7B, DAMO-NLP-SG/VideoLLaMA3-7B.

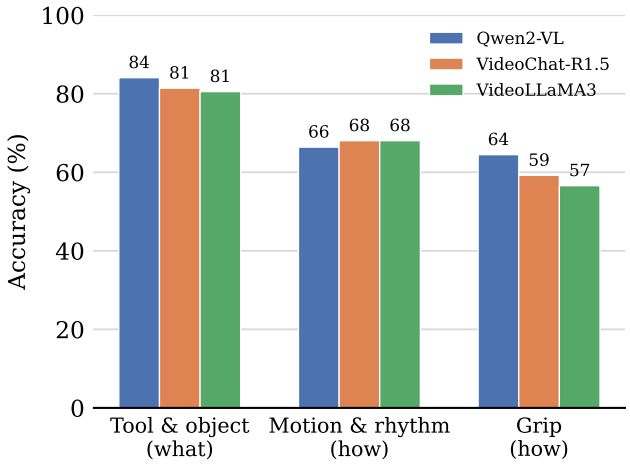


Fig. 2. Accuracy by kind of claim. All three models are strongest on tool and object claims (*what*) and drop on motion, rhythm, and grip claims (*how*).

A. *Reading* what *but not* how

Fig. 2 shows accuracy for each kind of attribute a claim is about. The pattern is the same for all three models: they are strongest on *what* is in the scene, the tools and objects (81–84%), weaker on the motion and rhythm (66–68%), and weakest on the grip (57–64%). Accuracy drops as the attribute moves from *what* is present toward *how* the action is done. This is our central result: the models read the visible tools and objects reliably, but the finer motion and grip cues that actually separate one style of execution from another are where they struggle.

B. *Confirming instead of checking*

The second pattern is just as strong, and it explains much of why the overall numbers stay near the baseline. As Fig. 3 shows, the models behave very differently depending on whether the attribute is actually there. When a claim is true of the clip, all three verify it well (81–85%), but when it is false, accuracy drops to 43–55%. The models tend to agree with a claim that sounds right rather than check whether it holds in the video. The errors show this clearly: of the 38 claims that all three models got wrong, 31 were cases where the attribute was absent but every model still answered "yes." In a robot setting this is a dangerous way to fail, acting on a style that was never demonstrated instead of simply missing one. Table II crosses the two effects. The drop on absent claims appears for both kinds of claim, but it is mild on *what* (58–74%) and severe on *how* (26–45%), so the two weaknesses compound rather than explain one another.

C. *Variation across actions and countries*

Accuracy also depends on how visually distinct an action is. Cutting, with its clear tool and clean separation of the object, is read most reliably (91–96%), while pouring is the hardest (61–67%), with cleaning, washing, and painting in between. A geographic gap appears as well: clips from Japan and the USA score highest, while India and Saudi Arabia score lowest. Uneven performance across locations is not surprising, but our per-country counts are small, so we read these differences as suggestive rather than conclusive.

TABLE II

ACCURACY (%) BY KIND OF CLAIM CROSSED WITH ITS LABEL. THE BIAS TOWARD CONFIRMING CLAIMS IS PRESENT IN BOTH KINDS BUT FAR STRONGER ON *how*.

	What		*How*	
Model	Present	Absent	Present	Absent
Qwen2-VL	89.3	73.7	76.7	44.9
VideoChat-R1.5	85.3	73.7	85.3	26.1
VideoLLaMA3	92.0	57.9	79.1	34.8
n claims	75	38	129	69

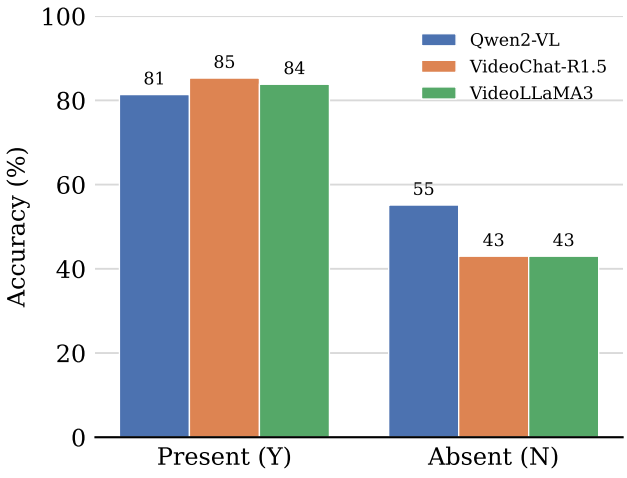


Fig. 3. Accuracy on present (Y) versus absent (N) claims. Every model is far weaker when the attribute is absent, revealing a bias toward confirming claims.

V. CONCLUSION AND DISCUSSION

We set out to test whether current VLMs can tell how an everyday action is performed, not just what is happening, and the answer is a qualified no. Across three models and 311 verified claims, they read the visible tools and objects reliably but do much worse on the motion, rhythm, and grip cues that set one style of execution apart from another. They also lean toward agreeing with a claim rather than checking it against the video, so their most common mistake is confirming a style the clip never shows. These two weaknesses compound, and they point at the same blind spot: how a task is actually carried out. This gap matters for robots that act around people. Teaching robots to imitate human video increasingly depends on models that can read how an action is done, not just what is done, so a VLM that misreads the *how*, or confirms it on demand, would weaken such methods before they even begin. Social navigation needs the same reading: whether a person is walking quickly or hesitating, waving a robot through or reaching past it. A model that confirms the manner its

instruction describes would have the robot yield to a gesture that was never made.

The study has clear limitations. The ground truth rests mostly on one labeler with a partial second pass that covered two of the five actions, the per-country counts are too small to treat as more than suggestive, and we cover three models on a single dataset. Even so, the pattern holds across all three, and it marks a concrete target for the next generation of video-language models, to read manner as closely as they already read content.

Future work could explore further probing of model understanding of "how", including by describing/explaining the claims in more concrete visual detail, as well as providing a baseline that does not take the visual information as input, to provide a reference point for the confirmation bias. Further models can be tested, and additional annotators added.

Acknowledgment

This work was partly funded by National Science Foundation award 2046853. We thank the University of Pittsburgh NOUR program for supporting this research. We also thank the IROS HRD and Social Robot Navigation workshop reviewers for suggestions for additional experiments.

References

[1] D. Moltisanti, F. Keller, H. Bilen, and L. Sevilla-Lara, "Learning action changes by measuring verb-adverb textual relationships," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 110–23 118.

[2] H. Doughty and C. G. Snoek, "How do you do it? fine-grained action understanding with pseudo-adverbs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 832–13 842.

[3] K. Grauman, A. Westbury, E. Byrne, V. Cartillier, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, D. Kukreja *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," *IEEE transactions on pattern analysis and machine intelligence*, 2024.

[4] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[5] Z. Yan, X. Li, Y. He, Z. Yue, X. Zeng, Y. Wang, Y. Qiao, L. Wang, and Y. Wang, "Videochat-r1.5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception," *arXiv preprint arXiv:2509.21100*, 2025.

[6] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li, P. Jin, W. Zhang, F. Wang, L. Bing, and D. Zhao, "Videollama 3: Frontier multimodal foundation models for image and video understanding," *arXiv preprint arXiv:2501.13106*, 2025.

[7] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014.

[8] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299–6308.

[9] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6202–6211.

[10] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[11] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022.

[12] C. Plizzari, T. Perrett, B. Caputo, and D. Damen, "What can a cook in italy teach a mechanic in india? action recognition generalisation over scenarios and locations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 13 656–13 666.

[13] A. Rai and A. Kovashka, "Generalizing sports feedback generation by watching competitions and reading books: A rock climbing case study," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.