

SemanticMoments: Training-Free Motion Similarity via Third Moment Features

Supplementary Material

1. Datasets

This supplementary document provides additional details about the SimMotion-Synthetic and SimMotion-Real benchmarks. As noted in the main text, code and data are available at <https://github.com/saarhub/semantic-moments>. Because motion similarity cannot be conveyed through static frames, we include qualitative video examples through two interactive webpages: `SimMotion-Synthetic.html` and `SimMotion-Real.html`, which are provided in the `motion_similarity_videos` directory of the supplementary material. These pages present representative triplets illustrating the structure and intended motion relationships within each benchmark.

SimMotion-Synthetic. The supplementary ZIP includes the system prompt and instruction template used to generate the four textual prompts in our pipeline: the base scene prompt, appearance-modification prompt, motion prompt, and negative-motion prompt. These templates clarify how the synthetic dataset was produced. SimMotion-Synthetic is designed to isolate motion from appearance by holding motion fixed while systematically varying non-motion factors such as identity, clothing, viewpoint, and rendering style. The accompanying webpage presents representative triplets: *reference*, *motion-preserving positive*, and *appearance-similar hard negative*, for each of the five variation categories, allowing readers to inspect the intended structure of the benchmark.

SimMotion-Real. For the real-world benchmark, Annotators were presented with a reference video and two candidate comparison videos and were asked to select which candidate exhibited more similar motion while ignoring appearance, background, and style. Each such comparison was rated independently by four annotators, and we utilized only responses with at least 75% agreement (three out of four votes) to ensure reliable motion-similarity judgments.

2. Additional Experimental Results

2.1. Gesture-Level Evaluation on Jester

We further extend our evaluation to the publicly available Jester gesture benchmark [1], which contains videos annotated with distinct gesture motion categories.

We evaluate whether SemanticMoments improves gesture-level separability in the embedding space under different video representations. To quantify this effect without

Table 1. **Gesture classification on Jester benchmark.** Top-1 majority vote and weighted kNN accuracy on the Jester validation set ($K=20$). *SemanticMoments* consistently improves performance across different backbones.

Method	Acc@1 _{maj}	Acc@1 _{w-kNN}	Acc@5 _{w-kNN}
X-CLIP	26.0	20.2	42.2
Clip4CLIP	8.5	8.3	11.3
TimeSFormer	16.7	10.3	20.1
SlowFast	19.7	17.2	34.8
VideoMoCo	12.1	12.0	24.4
I3D	26.8	25.8	53.2
VideoMAE	23.8	22.7	43.4
V-JEPA2	12.5	12.3	20.7
DINOv2	8.5	7.8	9.7
SemanticMoments _{DINO}	25.7	25.0	50.8
SemanticMoments _{VideoMAE}	<u>26.9</u>	<u>26.5</u>	55.0
SemanticMoments _{V-JEPA2}	28.6	28.3	47.2

training an additional classifier, we adopt a standard kNN evaluation protocol ($K=20$) on the validation set. For each query video, we retrieve its K nearest neighbors and predict the gesture label using their annotations. Majority-vote accuracy assigns the most frequent neighbor label, while weighted kNN additionally weights each neighbor contribution by its similarity to the query.

As shown in Table 1, applying SemanticMoments consistently improves the metrics across all backbones, indicating stronger motion representations.

2.2. Qualitative Evaluation on Kinetics-400

Although widely used for action recognition, Kinetics-400 is fundamentally misaligned with the goal of motion-centric video similarity. To illustrate this, we include an additional webpage, `Kinetics.html`, located in the `motion_similarity_videos` directory of the supplementary material. This page presents qualitative examples from the Kinetics-400 test set. We focus on cases where the nearest neighbor retrieved by our SemanticMoments (DINO) representation exhibits semantically similar motion to the query but belongs to a different Kinetics label. For each such query, we also show several randomly selected videos from the query’s own class. These examples reveal that same-label videos often display substantially different motions, while semantically similar motions may come from unrelated classes. This highlights the strong appearance bias of Kinetics and explains why it is not used as a benchmark in the main paper.

References

- [1] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale

video dataset of human gestures. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. [1](#)