MGSampler: An Explainable Sampling Strategy for Video Action Recognition **Supplementary Material**

Yuan Zhi   Zhan Tong   Limin Wang✉   Gangshan Wu
State Key Laboratory for Novel Software Technology, Nanjing University, China
{yuanzhi,tongzhan}@smail.nju.edu.cn, {lmwang,gswu}@nju.edu.cn

A. Evaluation on multiple views

The goal of MGSampler is to provide a holistic sparse sampler and only sample one clip from each video for efficient inference. That is a widely used testing scheme by recent methods in Sth-Sth dataset [3]. Indeed, multi-view testing could further improve the performance but also increase computaitonal cost. We perform multi-view testing (2 clips and 3 crops) on our MGSampler in the same manner with TSM[6], and the result is shown in Table 1.

Model	Frames	Test-Views	Sampler	Top-1 Acc
TSM-R50	8	1×1	TSN	57.9
TSM-R50	8	1×1	**MG**	59.8(**+1.9**)
TSM-R50	8	2×3	TSN	61.2
TSM-R50	8	2×3	**MG**	62.9(**+1.7**)

Table 1. Multi-view testing on **Something-Something V2.**

B. Use MGSampler as a clip sampler

Our MGSampler could be easily adapted to dense clip sampling. The original dense methods samples 8 frames from continuous 32 frames with stride 4. Our MGSampler can adaptively sample a 8-frame clip guided by accumulation curve from the same continuous 32 frames. The results on Sth-Sth V2 are shown in Table 2, which demonstrates the effectiveness of MGSampler on dense sampling.

Model	Frames	Test-Views	Clip Sampler	Top-1 Acc
SlowOnly-R50	8	1×1	fixed stride	57.7
SlowOnly-R50	8	1×1	**MG**	58.5(**+0.8**)
SlowOnly-R50	8	10×3	fixed stride	62.1
SlowOnly-R50	8	10×3	**MG**	62.5(**+0.4**)

Table 2. MGSampler extension as a dense clip sampler. Testing with SlowOnly-R50 [2] on **Something-Something V2.**

C. Results on untrimmed videos

we extend MGSampler to untrimmed video testing. The results in ActivityNet [1] is reported in Table 3. We first perform sparse frame sampling in a TSN-like framework, and our MGSampler is better than TSN by 1.4%. Then we use MGSampler to perform dense clip sampling as in **Section B** and it is better than standard dense clip sampling by 0.7%.

Model	Frames	Test-Views	Sampler	Top-1 Acc
SlowOnly-R50	8	1×1	TSN	77.4
SlowOnly-R50	8	1×1	**MG**	78.8(**+1.4**)
SlowOnly-R50	8	10×3	8×8 clip	80.3
SlowOnly-R50	8	10×3	**MG**(clip)	81.0(**+0.7**)

Table 3. Performance comparison on **ActivityNet 1.3.**

D. Visualization analysis

More examples of comparison between uniform sample and motion-guided sample on Sth-Sth [3], Diving48 [5], UCF101 [8], HMDB [4], Jester [7] datasets. The left column of Figure 1 is the cumulative distribution motion and the right column is the sampled frames.

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[2] 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.

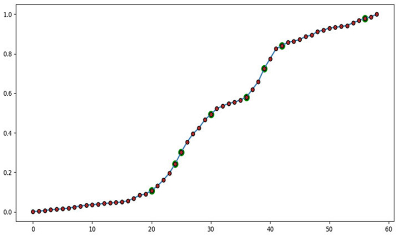
[3] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *ICCV*, 2017.

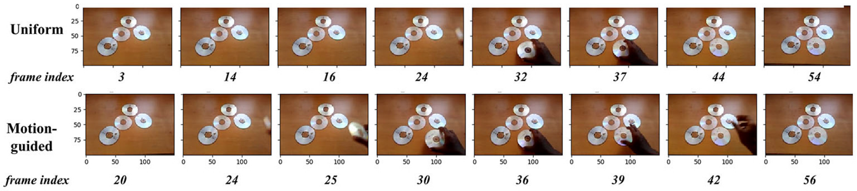
[4] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: A Large Video Database for Human Motion Recognition. In *ICCV*, 2011.

[5] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *ECCV*, pages 513–528, 2018.

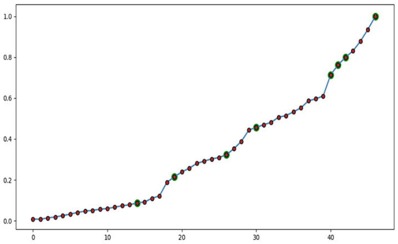
✉: Corresponding author.

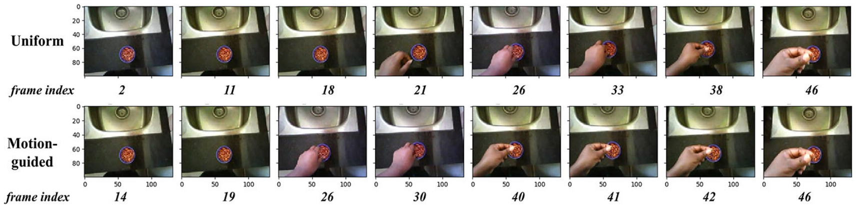
Something-Something



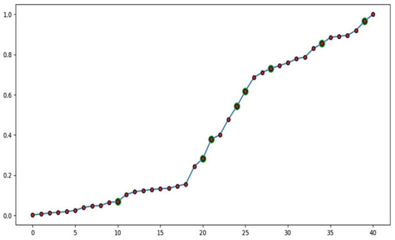


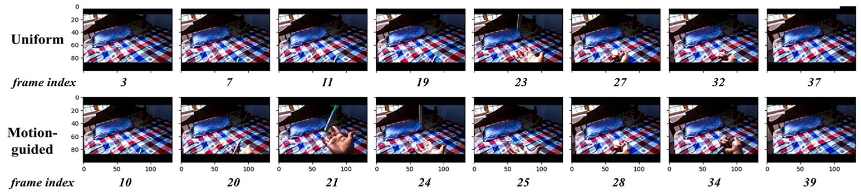
label: putting something similar to other things that are already on the table





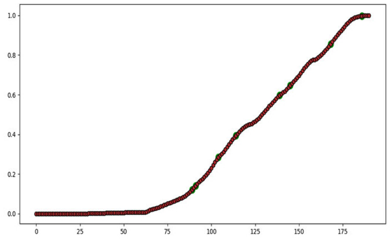
label: taking something out of something

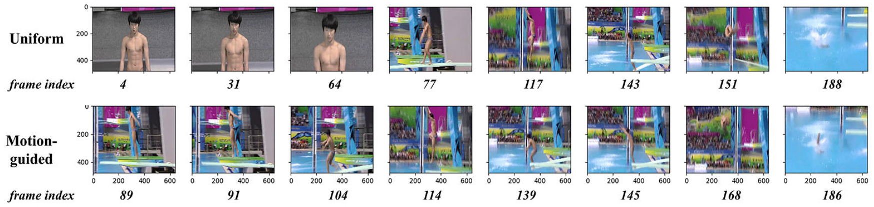




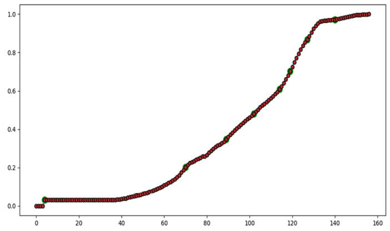
label: throwing something in the air and catch it

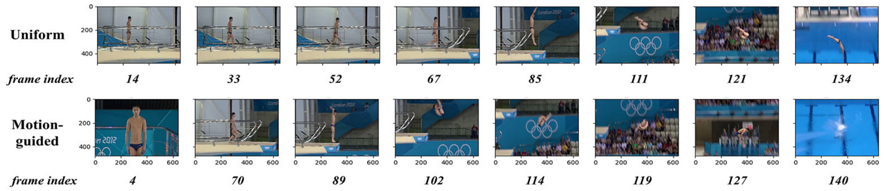
Diving48



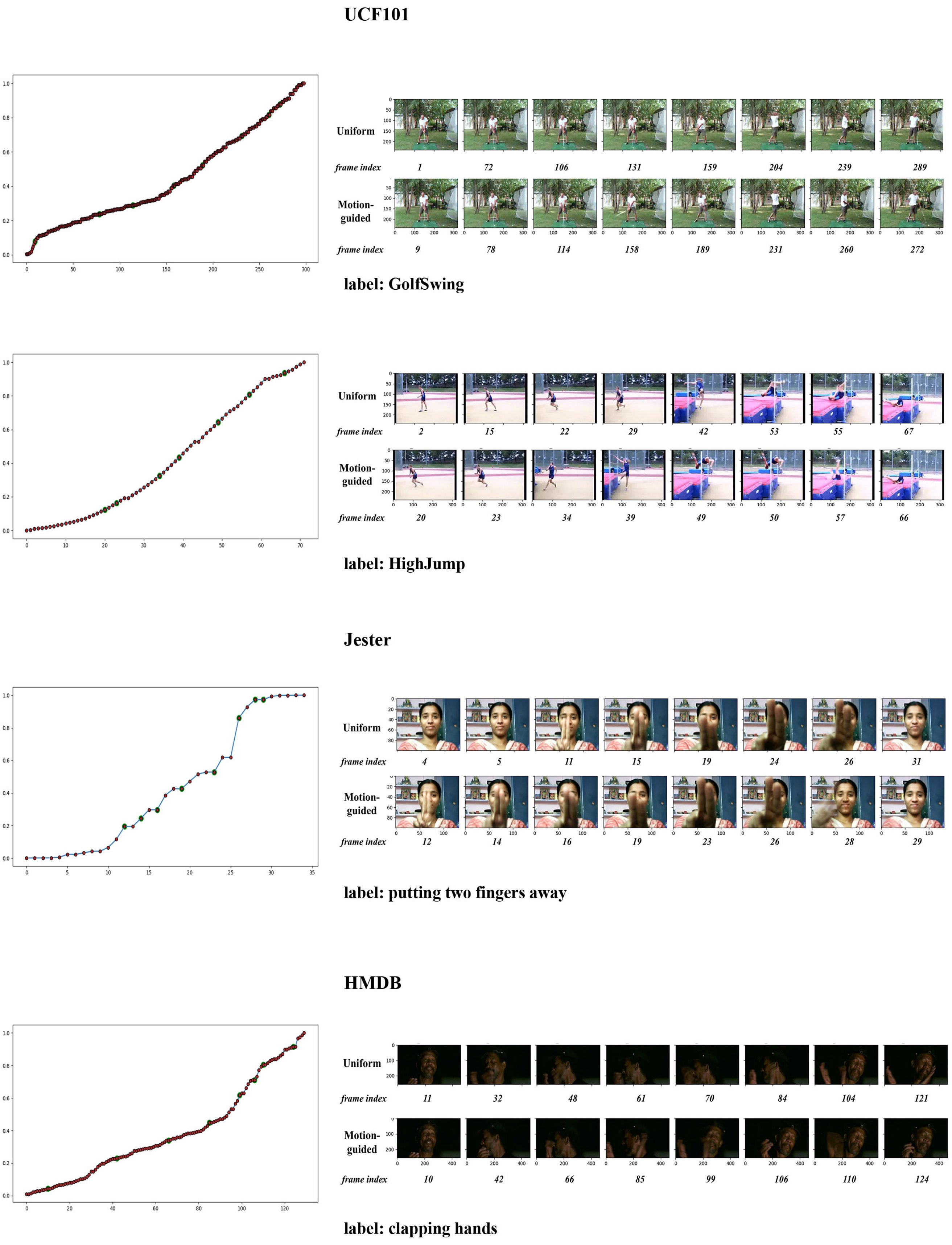


label: [“Forward", "35som", "NoTwis", "PIKE"]





label: ["Forward", "25som", "NoTwis", "TUCK"]



UCF101

label: GolfSwing

label: HighJump

Jester

label: putting two fingers away

HMDB

label: clapping hands

Figure 1. Examples of comparison between uniform sample and motion-guided sample on five datasets.

[6] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019.

[7] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *ICCVW)*, pages 2874–2882, 2019.

[8] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv preprint arXiv:1212.0402*, 2012.