Unbiasing through Textual Descriptions: Mitigating Representation Bias in Video Benchmarks. Supplementary Material

Nina Shvetsova[1,2,3]*, Arsha Nagrani[4], Bernt Schiele[3], Hilde Kuehne[1,2,5], Christian Rupprecht[4]
[1] Goethe University Frankfurt, [2] Tuebingen AI Center/University of Tuebingen, [3] MPI for Informatics, Saarland Informatics Campus,
[4] University of Oxford, [5] MIT-IBM Watson AI Lab
shvetsov@uni-frankfurt.de

In the supplementary material, we provide additional experimental results, implementation details, and qualitative examples. Furthermore, we discuss the UTD dataset's license, limitations, and broader impact, and provide a datasheet [9] for the UTD dataset. Specifically, we first present additional experimental results, including extended results on benchmarking video models and common-sense bias, ablation studies, as well as analysis of class distribution in UTD-splits in Appendix A. Next, we provide further implementation details of our UTD method in Appendix B. We then demonstrate qualitative examples from our UTD-descriptions and UTD-splits datasets, along with samples from the user study, in Appendix C. Finally, we discuss the UTD dataset's license in Appendix D, the limitations of our work and its broader impact in Appendix E, and provide a datasheet for the UTD dataset in Appendix F.

A. Additional Results

A.1. Benchmarking Video Models

In this section, we present additional benchmarking results for state-of-the-art video models on our object-debiased UTD-splits. Specifically, we extend the analysis presented in Tables 5 and 6 of the main paper by including three considered classification datasets, namely Kinetics 600, Kinetics 700, and MiT, and three considered retrieval datasets, namely LSMDC, YouCook2, and Spoken-MiT. In Tab. A.1, we provide the performance of selected video models on all classification datasets, evaluated both on the full test/val sets and on our debiased UTD-splits. The models were chosen based on their strong performance in Table 5 of the main paper. And in Tab. A.2, we present the evaluation results for video models across all considered retrieval datasets.

A.2. Common Sense Bias

In Tab. A.3, we present additional common sense bias results for all 16 conceptual-temporal combinations across the 12 datasets considered. The observed effects align with results discussed in the main paper. Specifically, the overall classification performance drops only slightly when predictions are based solely on objects compared to the objects+composition+activities setup across most datasets.

A.3. Ablation Study

As discussed in the main paper, to estimate representation bias, we design a strong model that performs action classification and text-to-video retrieval based solely on the textual descriptions of videos. Throughout our pipeline, we utilize state-of-the-art models, namely LLaVA-1.6-Mistral-7B [21] as the VLM, Mistral-7B-Instruct-v0.2 [12] as the LLM, and SFR-Embedding-Mistral [23] for text embedding model. In this section, we provide additional analysis of our model.

Text Embedding Model. First, in Tab. A.4, we compare four strong models for text encoding (χ in Figure 3 of the main paper). Namely, we consider large versions of the CLIP [28] text encoder and LongCLIP [40] text encoder, which extends CLIP to better handle long-text inputs. Additionally, we examine two LLM-based text embedding models: E5-Mistral-7B-Instruct [35] and its fine-tuned version, SFR-Embedding-Mistral [23], trained on more data. As shown in Tab. A.4, the text embedding models E5-Mistral-7B-Instruct and SFR-Embedding-Mistral, both pretrained on large text datasets to effectively encode text for tasks such as information retrieval, outperform the CLIP-based text embedding models. In our pipeline, we employ the best-performing SFR-Embedding-Mistral model.

Vision-Language Model. Next, in A.5, we ablate two VLM models for extracting detailed textual descriptions.

*The work was done during PhD visit to the University of Oxford within the ELLIS PhD program.

	UCF	UCF-UTD-split	UCF-UTD-s. balanced	SSv2	SSv2-UTD-split	SSv2-UTD-s. balanced	K400	K400-UTD-split	K400-UTD-s. balanced	K600	K600-UTD-split	K600-UTD-s. balanced	K700	K700-UTD-split	K700-UTD-s. balanced	MiT	MiT-UTD-split	MiT-UTD-s. balanced
videomae-L-UH	95.7	$90.9_{-4.8}$	$91.0_{-4.7}$	67.3	$65.4_{-1.9}$	$66.2_{-1.1}$	75.8	$59.3_{-16.5}$	$63.3_{-12.5}$	76.9	$61.1_{-15.8}$	$65.8_{-11.1}$	65.6	$47.6_{-18.0}$	$52.7_{-12.9}$	38.1	$26.4_{-11.7}$	$30.8_{-7.3}$
videomaev2-B-K710-fnK710	99.0	$98.1_{-0.9}$	$97.6_{-1.4}$	57.1	$54.7_{-2.4}$	$55.9_{-1.2}$	85.0	$71.9_{-13.1}$	$75.5_{-9.5}$	85.6	$73.2_{-12.4}$	$77.2_{-8.4}$	75.9	$60.0_{-15.9}$	$65.3_{-10.6}$	40.9	$28.9_{-12.0}$	$33.9_{-7.0}$
allinone-B-WV2M	84.5	$63.1_{-21.4}$	$72.9_{-11.6}$	26.2	$22.5_{-3.7}$	$24.7_{-1.5}$	66.9	$42.6_{-24.3}$	$50.8_{-16.1}$	68.0	$45.0_{-23.0}$	$53.3_{-14.7}$	55.2	$32.1_{-23.1}$	$40.0_{-15.2}$	29.9	$16.9_{-13.0}$	$22.0_{-7.9}$
umt-B-fnK710	99.0	$96.8_{-2.2}$	$97.1_{-1.9}$	49.4	$46.6_{-2.8}$	$48.1_{-1.3}$	85.6	$72.4_{-13.2}$	$76.2_{-9.4}$	86.2	$73.8_{-12.4}$	$78.1_{-8.1}$	76.9	$61.5_{-15.4}$	$66.8_{-10.1}$	39.7	$27.1_{-12.6}$	$32.1_{-7.6}$
umt-L-fnK710	98.9	$96.7_{-2.2}$	$97.1_{-1.8}$	57.9	$55.5_{-2.4}$	$56.7_{-1.2}$	89.0	$78.7_{-10.3}$	$81.6_{-7.4}$	89.0	$78.9_{-10.1}$	$82.3_{-6.7}$	82.1	$69.3_{-12.8}$	$74.0_{-8.1}$	44.5	$32.3_{-12.2}$	$37.2_{-7.3}$
videomamba-vm-25M	94.3	$83.0_{-11.3}$	$86.5_{-7.8}$	48.7	$45.9_{-2.8}$	$47.1_{-1.6}$	78.4	$59.1_{-19.3}$	$64.6_{-13.8}$	78.5	$60.2_{-18.3}$	$66.6_{-11.9}$	68.1	$47.5_{-20.6}$	$54.5_{-13.6}$	37.9	$24.5_{-13.4}$	$29.7_{-8.2}$
internvid-B-10M-FLT	94.0	$81.1_{-12.9}$	$84.7_{-9.3}$	48.1	$45.2_{-2.9}$	$46.7_{-1.4}$	78.6	$59.3_{-19.3}$	$64.9_{-13.7}$	78.7	$60.1_{-18.6}$	$66.6_{-12.1}$	68.5	$47.8_{-20.7}$	$55.0_{-13.5}$	39.3	$26.0_{-13.3}$	$31.1_{-8.2}$
internvid-B-200M	94.5	$82.9_{-11.6}$	$85.2_{-9.3}$	54.4	$51.8_{-2.6}$	$52.8_{-1.6}$	80.0	$61.7_{-18.3}$	$66.8_{-13.2}$	79.9	$62.0_{-17.9}$	$68.5_{-11.4}$	70.2	$50.3_{-19.9}$	$57.6_{-12.6}$	39.9	$26.6_{-13.3}$	$31.9_{-8.0}$
internvid-L-10M-FLT	95.5	$86.3_{-9.2}$	$88.4_{-7.1}$	53.6	$50.9_{-2.7}$	$52.1_{-1.5}$	81.9	$64.7_{-17.2}$	$69.6_{-12.3}$	81.5	$64.8_{-16.7}$	$70.8_{-10.7}$	72.4	$53.3_{-19.1}$	$60.3_{-12.1}$	41.9	$28.9_{-13.0}$	$33.8_{-8.1}$
internvid-L-200M	96.8	$90.5_{-6.3}$	$91.9_{-4.9}$	63.1	$61.0_{-2.1}$	$61.9_{-1.2}$	84.6	$69.4_{-15.2}$	$73.9_{-10.7}$	84.3	$69.3_{-15.0}$	$75.1_{-9.2}$	76.4	$59.4_{-17.0}$	$66.0_{-10.4}$	44.3	$31.7_{-12.6}$	$36.9_{-7.4}$

Table A.1. **Benchmarking video models in action classification on all six considered classification datasets.** We report accuracy on full test/val sets and our debiased UTD-splits. The accuracy differences with respect to the full test/val sets are color-coded.

	MSR	MSR-UTD-split	DDM	DDM-UTD-split	ANet	ANet-UTD-split	LSMDC	LSMDC-UTD-split	YC2	YC2-UTD-split	S-MiT	S-MiT-UTD-split
umt-b-5M	30.0	$20.5_{-9.5}$	30.2	$20.6_{-9.6}$	28.6	$22.3_{-6.3}$	14.1	$9.2_{-4.9}$	6.1	$5.2_{-0.9}$	47.9	$34.0_{-13.9}$
umt-b-17M	35.6	$26.3_{-9.3}$	37.7	$27.6_{-10.1}$	34.2	$27.3_{-6.9}$	16.6	$11.1_{-5.5}$	8.4	$6.9_{-1.5}$	53.5	$39.5_{-14.0}$
umt-b-25M	35.3	$24.8_{-10.5}$	34.2	$24.6_{-9.6}$	25.1	$19.7_{-5.4}$	13.1	$9.0_{-4.1}$	10.3	$8.7_{-1.6}$	53.9	$40.2_{-13.7}$
umt-l-5M	34.8	$24.9_{-9.9}$	33.5	$21.8_{-11.7}$	34.8	$28.7_{-6.1}$	21.5	$16.5_{-5.0}$	7.1	$5.8_{-1.3}$	51.9	$37.9_{-14.0}$
umt-l-17M	43.6	$31.1_{-12.5}$	46.3	$35.6_{-10.7}$	45.9	$38.7_{-7.2}$	21.6	$16.7_{-4.9}$	14.4	$11.9_{-2.5}$	60.7	$46.9_{-13.8}$
umt-l-25M	42.3	$30.6_{-11.7}$	43.6	$33.5_{-10.1}$	36.7	$30.4_{-6.3}$	19.4	$14.2_{-5.2}$	15.1	$12.8_{-2.3}$	60.8	$47.9_{-12.9}$
videomamba-vm-5M	33.3	$23.0_{-10.3}$	37.1	$27.1_{-10.0}$	37.1	$30.1_{-7.0}$	17.6	$12.7_{-4.9}$	6.5	$5.6_{-0.9}$	47.6	$34.0_{-13.6}$
videomamba-vm-17M	34.9	$25.5_{-9.4}$	40.6	$28.9_{-11.7}$	40.4	$33.0_{-7.4}$	20.1	$15.5_{-4.6}$	7.7	$6.6_{-1.1}$	51.6	$38.3_{-13.3}$
videomamba-vm-25M	34.9	$25.5_{-9.4}$	41.4	$30.5_{-10.9}$	41.1	$33.8_{-7.3}$	20.4	$15.4_{-5.0}$	9.3	$7.9_{-1.4}$	53.2	$39.7_{-13.5}$
internvid-B-10M-FLT	37.9	$25.4_{-12.5}$	28.6	$17.2_{-11.4}$	24.4	$18.8_{-5.6}$	17.0	$10.7_{-6.3}$	8.1	$5.9_{-2.2}$	48.9	$34.6_{-14.3}$
internvid-B-200M	38.1	$24.7_{-13.4}$	30.2	$19.4_{-10.8}$	26.2	$20.1_{-6.1}$	18.3	$11.9_{-6.4}$	8.6	$6.4_{-2.2}$	49.8	$35.8_{-14.0}$
internvid-L-10M	26.7	$18.7_{-8.0}$	22.6	$15.6_{-7.0}$	21.5	$16.3_{-5.2}$	11.4	$7.0_{-4.4}$	6.8	$5.5_{-1.3}$	35.9	$23.6_{-12.3}$
internvid-L-WV10M	26.5	$17.6_{-8.9}$	22.2	$13.8_{-8.4}$	23.1	$17.5_{-5.6}$	12.3	$7.6_{-4.7}$	6.7	$5.6_{-1.1}$	38.8	$25.9_{-12.9}$
internvid-L-10M-DIV	37.3	$24.2_{-13.1}$	26.9	$15.9_{-11.0}$	23.2	$17.7_{-5.5}$	15.7	$10.3_{-5.4}$	8.6	$6.7_{-1.9}$	48.9	$34.4_{-14.5}$
internvid-L-10M-FLT	38.7	$26.3_{-12.4}$	29.2	$18.8_{-10.4}$	24.5	$19.0_{-5.5}$	19.5	$13.5_{-6.0}$	9.4	$7.5_{-1.9}$	50.5	$36.2_{-14.3}$
internvid-L-50M	32.4	$22.0_{-10.4}$	26.5	$18.3_{-8.2}$	24.4	$18.0_{-6.4}$	17.8	$11.7_{-6.1}$	8.0	$6.4_{-1.6}$	45.7	$31.5_{-14.2}$
internvid-L-200M	38.2	$24.8_{-13.4}$	30.3	$20.2_{-10.1}$	28.7	$22.1_{-6.6}$	20.1	$13.7_{-6.4}$	11.0	$8.8_{-2.2}$	53.7	$39.0_{-14.7}$

Table A.2. **Benchmarking video-language models in text-to-video retrieval on all six considered retrieval datasets.** We report accuracy on full test/val sets and our debiased UTD-splits. The accuracy differences with respect to the full test/val sets are color-coded.

Specifically, we evaluate LLaVA-v1.5-7B [20] and LLaVA-1.6-Mistral-7B [21], finding that the latter model achieves the best performance.

Comparison to CLIP. Finally, we evaluate how well our model, which relies solely on textual descriptions, performs for zero-shot video classification and retrieval compared to the CLIP baseline [28] The results in Tab. A.6 demonstrate that our model generally outperforms CLIP ViT-B/32 and performs almost on par with the CLIP ViT-L/14 backbone. This highlights the feasibility of performing action classification and video retrieval based purely on textual descriptions. We further observe that caption-based retrieval performs better compared to caption-based action classification. We attribute this to the fact that captions capture more specific aspects of individual videos, which are also reflected in the generated captions.

A.4. Class Distribution in UTD-Splits

In Fig. A.1, we show the class distribution in our UTD- and UTD-balanced test/validation splits, in comparison with the original class distribution in the full test/validation splits. We observe that the results align well with our expectations, for example, the class "Apply Eye Makeup" in UCF-101 is significantly reduced in the UTD-split due to a strong object bias.

B. Additional Implementation Details

In this section, we provide further details about our UTD method.

B.1. Obtaining Textual Descriptions

We used a few-shot in-context learning strategy [3] when prompting LLM to extract objects, activities, and verbs from objects+composition+activities descriptions. Namely, we used 3-shots for objects and activities and 5-shots for verbs. We also did simple postprocessing of LLM output. Since the LLM is prompted to output an enumerated list (see prompts in Tab. F.1), we delete numeration and delete any output in brackets.

We list the prompts used in our work to obtain textual de-

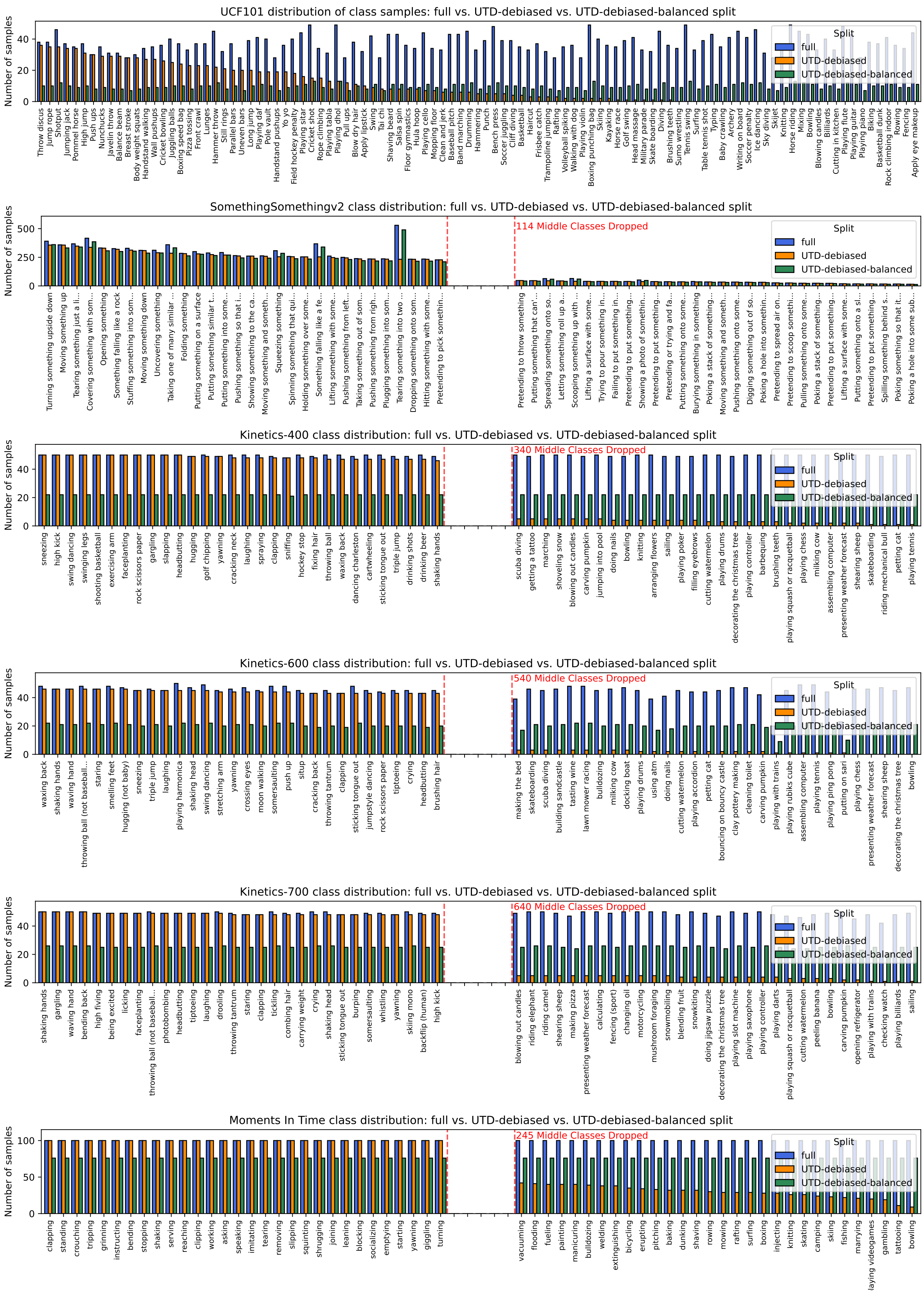


Figure A.1. Comparison of class distribution in full test/val split vs. UTD-debiased split vs. UTD-debiased-balanced split for six considered classification datasets.

Action Classification Datasets

	UCF seq.-of-f.	UCF avg.-over-f.	UCF max-score-f.	UCF middle f.	SSv2 seq.-of-f.	SSv2 avg.-over-f.	SSv2 max-score-f.	SSv2 middle f.	K400 seq.-of-f.	K400 avg.-over-f.	K400 max-score-f.	K400 middle f.
obj+comp+act	66.3	$66.7_{+0.4}$	$66.5_{+0.2}$	$61.3_{-5.0}$	6.4	$6.8_{+0.4}$	$7.4_{+1.0}$	$6.0_{-0.4}$	48.0	$46.6_{-1.4}$	48.0	$39.7_{-8.3}$
objects	$63.3_{-3.0}$	$61.9_{-4.4}$	$62.5_{-3.8}$	$57.4_{-8.9}$	$5.3_{-1.1}$	$4.6_{-1.8}$	$5.1_{-1.3}$	$4.0_{-2.4}$	$45.9_{-2.1}$	$41.0_{-7.0}$	$42.3_{-5.7}$	$35.0_{-13.0}$
activities	$67.4_{+1.1}$	$67.4_{+1.1}$	$65.7_{-0.6}$	$59.0_{-7.3}$	6.4	$6.0_{-0.4}$	$6.0_{-0.4}$	$4.9_{-1.5}$	$45.2_{-2.8}$	$42.1_{-5.9}$	$42.8_{-5.2}$	$31.3_{-16.7}$
verbs	$50.8_{-15.5}$	$41.5_{-24.8}$	$43.7_{-22.6}$	$31.1_{-35.2}$	$5.8_{-0.6}$	$4.4_{-2.0}$	$4.1_{-2.3}$	$3.1_{-3.3}$	$24.8_{-23.2}$	$16.1_{-31.9}$	$20.1_{-27.9}$	$10.9_{-37.1}$

	K600 seq.-of-f.	K600 avg.-over-f.	K600 max-score-f.	K600 middle f.	K700 seq.-of-f.	K700 avg.-over-f.	K700 max-score-f.	K700 middle f.	MiT seq.-of-f.	MiT avg.-over-f.	MiT max-score-f.	MiT middle f.
obj+comp+act	44.1	$42.1_{-2.0}$	$43.3_{-0.8}$	$35.6_{-8.5}$	39.0	$37.1_{-1.9}$	$38.7_{-0.3}$	$31.1_{-7.9}$	22.6	$23.4_{+0.8}$	$21.4_{-1.2}$	$20.1_{-2.5}$
objects	$41.8_{-2.3}$	$36.5_{-7.6}$	$37.5_{-6.6}$	$31.0_{-13.1}$	$37.0_{-2.0}$	$32.2_{-6.8}$	$33.3_{-5.7}$	$26.7_{-12.3}$	$21.0_{-1.6}$	$19.4_{-3.2}$	$20.1_{-2.5}$	$17.6_{-5.0}$
activities	$41.4_{-2.7}$	$38.1_{-6.0}$	$38.5_{-5.6}$	$28.1_{-16.0}$	$36.7_{-2.3}$	$33.0_{-6.0}$	$33.8_{-5.2}$	$24.0_{-15.0}$	$21.0_{-1.6}$	$20.0_{-2.6}$	$20.1_{-2.5}$	$15.6_{-7.0}$
verbs	$21.4_{-22.7}$	$13.8_{-30.3}$	$16.7_{-27.4}$	$8.9_{-35.2}$	$17.6_{-21.4}$	$11.0_{-28.0}$	$13.6_{-25.4}$	$7.0_{-32.0}$	$16.2_{-6.4}$	$12.2_{-10.4}$	$14.1_{-8.5}$	$9.2_{-13.4}$

Text-to-Video Retrieval Datasets

	MSR seq.-of-f.	MSR avg.-over-f.	MSR max-score-f.	MSR middle f.	DDM seq.-of-f.	DDM avg.-over-f.	DDM max-score-f.	DDM middle f.	ActN seq.-of-f.	ActN avg.-over-f.	ActN max-score-f.	ActN middle f.
obj+comp+act	36.6	$31.9_{-4.7}$	$33.4_{-3.2}$	$23.4_{-13.2}$	27.2	$26.5_{-0.7}$	$29.1_{+1.9}$	$19.8_{-7.4}$	26.6	26.6	$21.5_{-5.1}$	$13.5_{-13.1}$
objects	$32.1_{-4.5}$	$28.7_{-7.9}$	$29.7_{-6.9}$	$17.8_{-18.8}$	$27.0_{-0.2}$	$26.5_{-0.7}$	$25.5_{-1.7}$	$16.8_{-10.4}$	$24.8_{-1.8}$	$22.7_{-3.9}$	$17.8_{-8.8}$	$11.5_{-15.1}$
activities	$25.1_{-11.5}$	$22.4_{-14.2}$	$18.6_{-18.0}$	$11.6_{-25.0}$	$21.1_{-6.1}$	$21.0_{-6.2}$	$18.4_{-8.8}$	$11.3_{-15.9}$	$21.4_{-5.2}$	$17.8_{-8.8}$	$13.2_{-13.4}$	$8.4_{-18.2}$
verbs	$10.5_{-26.1}$	$8.7_{-27.9}$	$7.8_{-28.8}$	$4.2_{-32.4}$	$7.0_{-20.2}$	$6.0_{-21.2}$	$5.5_{-21.7}$	$3.5_{-23.7}$	$7.4_{-19.2}$	$5.1_{-21.5}$	$3.9_{-22.7}$	$2.5_{-24.1}$

	LSMDC seq.-of-f.	LSMDC avg.-over-f.	LSMDC max-score-f.	LSMDC middle f.	YC2 seq.-of-f.	YC2 avg.-over-f.	YC2 max-score-f.	YC2 middle f.	S-MiT seq.-of-f.	S-MiT avg.-over-f.	S-MiT max-score-f.	S-MiT middle f.
obj+comp+act	17.0	$16.4_{-0.6}$	$18.3_{+1.3}$	$12.7_{-4.3}$	8.4	$8.6_{+0.2}$	$8.9_{+0.5}$	$6.1_{-2.3}$	45.9	$43.8_{-2.1}$	$46.1_{+0.2}$	$35.5_{-10.4}$
objects	$13.6_{-3.4}$	$11.7_{-5.3}$	$12.3_{-4.7}$	$9.2_{-7.8}$	$7.9_{-0.5}$	$6.2_{-2.2}$	$7.1_{-1.3}$	$4.9_{-3.5}$	$29.8_{-16.1}$	$27.5_{-18.4}$	$26.4_{-19.5}$	$18.2_{-27.7}$
activities	$14.7_{-2.3}$	$13.6_{-3.4}$	$11.4_{-5.6}$	$7.8_{-9.2}$	$8.1_{-0.3}$	$7.3_{-1.1}$	$6.3_{-2.1}$	$4.1_{-4.3}$	$41.1_{-4.8}$	$38.4_{-7.5}$	$37.8_{-8.1}$	$29.3_{-16.6}$
verbs	$5.5_{-11.5}$	$6.0_{-11.0}$	$4.5_{-12.5}$	$3.0_{-14.0}$	$1.2_{-7.2}$	$1.1_{-7.3}$	$1.0_{-7.4}$	$0.6_{-7.8}$	$13.1_{-32.8}$	$9.5_{-36.4}$	$8.8_{-37.1}$	$4.8_{-41.1}$

Table A.3. **Evaluation of common sense bias with respect to all 16 conceptual-temporal combinations on all 12 considered datasets.** We color code with respect to the difference to objects+composition+activities (obj+comp+act) concepts in sequence-of-frames (seq.-of-f.) temporal setup.

Text Embedding Model	Action Classification UCF	Action Classification SSv2	Text-to-Video Retrieval MSR	Text-to-Video Retrieval DDM
CLIP [28] text encoder	51.1	1.8	6.1	7.0
Long-CLIP [40] text encoder	48.5	2.8	24.0	19.0
E5-Mistral-7B-Instruct [35]	65.6	6.2	35.1	26.7
SFR-Embedding-Mistral [23]	**66.3**	**6.4**	**36.6**	**27.2**

Table A.4. **Ablation on text embedding model.** Evaluation is performed in objects+composition+activities concepts in sequence-of-frames temporal setup on the full test/validation splits of the respective datasets. We report accuracy for classification and recall@1 for retrieval. Selected option is bolded.

scriptions of frames for different concept categories, namely objects+composition+activities ($d_{n,i}$), objects ($o_{n,i}$), activities ($a_{n,i}$), and verbs ($\nu_{n,i}$), in Tab. F.1

B.2. Getting Text Embeddings

Following [23], we used the prompt template "`Instruct : <instruction>\nQuery: <input text>`" to obtain text embeddings for various input text descriptions using the SFR-Embedding-Mistral model [23]. In Tab. F.2, we provide the instructions used for action classification in different settings, and in Tab. F.3, we present instructions used for text-to-video retrieval.

Vision-Language Model	Action Classification UCF	Action Classification SSv2	Text-to-Video Retrieval MSR	Text-to-Video Retrieval DDM
LLaVA-v1.5-7B [20]	61.8	5.9	35.5	**28.7**
LLaVA-1.6-Mistral-7B [21]	**66.3**	**6.4**	**36.6**	27.2

Table A.5. **Ablation on vision-language model.** Evaluation is performed in objects+composition+activities concepts in sequence-of-frames temporal setup on the full test/validation splits of the respective datasets. We report accuracy for classification and recall@1 for retrieval. Selected option is bolded.

B.3. Unbiasing Datasets

As stated in the main paper, we debias the test/val sets of the considered datasets by excluding samples that are classified or retrieved correctly ($M(v_n, \phi) = 1$) based on *object sequence-of-frame* representation.

For text-to-video retrieval, we debias datasets with re-

	Action Classification						Text-to-Video Retrieval					
	UCF	SSv2	K400	K600	K700	MiT	MSR	DDM	ANet	LSMDC	YC2	S-MiT
CLIP ViT-B/32	<u>67.7</u>	1.8	46.7	41.3	34.8	18.6	31.4	26.3	20.4	14.3	4.9	34.7
CLIP ViT-L/14	**75.9**	<u>2.8</u>	**58.7**	**54.7**	**48.6**	**23.7**	<u>36.3</u>	**29.6**	<u>26.2</u>	**19.9**	<u>8.1</u>	<u>40.4</u>
Ours (obj+comp+act, sequence-of-frames)	66.3	**6.4**	<u>48.0</u>	<u>44.1</u>	<u>39.0</u>	<u>22.6</u>	**36.6**	<u>27.2</u>	**26.6**	<u>17.0</u>	**8.4**	**45.9**

Table A.6. **Comparison in zero-shot action classification and text-to-video retrieval.** Note, that our model uses only textual descriptions of video. Evaluation is performed on full test/val splits of the respective datasets. We report accuracy for classification and Recall@1 for retrieval.

spect to *common-sense bias*. To minimize the impact of random fluctuations in the process, we prompt the text embedding models with three different prompts for query captions and three different prompts for object textual descriptions for videos, generating three embeddings for each query caption and for each video. Prompts are reported in Tab. F.4. Consequently, we perform zero-shot text-to-video retrieval inference using all nine combinations of query embeddings and video embeddings, resulting in nine predictions. We exclude a text query from the test/val set only if all nine predictions agree on the correct Top-1 retrieval for the corresponding video.

For action classification, we debias datasets with respect to *dataset bias*. Specifically, we generate three different embeddings for the object textual descriptions of each video using three different instructions for the text embedding model (Tab. F.4), resulting in three sets of video embeddings. Using each set of embeddings, we further train three linear classifiers on bootstrapped training sets (we sample training sets using sampling replacement to match the original training set size). A sample is excluded from the test/val set only if all nine models (across the three embedding sets and three linear models) agree on the correct Top-1 classification. The percentage of removed samples is determined automatically based on the extent of object representation bias in the dataset.

Since debiasing may disproportionately remove certain label classes in classification datasets, we additionally construct balanced UTD splits. While maintaining the total number of removed samples as in the original debiasing method, we adjust the number of samples removed from each class based on their average confidence (across nine models) to preserve the original class proportions.

Reliability of filtering biased samples. To summarize, we ensure robust and reliable filtering of biased samples by leveraging state-of-the-art VLMs and LLMs, applying prompt engineering, and employing an in-context learning strategy to extract specific concepts (e.g., objects) while minimizing leakage from unrelated concepts. We also conduct a user study to validate the reliability of the extracted concepts. To further mitigate false positives, we aggregate predictions across nine different prompt/model combinations.

B.4. Benchmarking Video Models

Table B.1 summarizes all 30 video models evaluated in this work, detailing their architectural backbones, pretraining datasets, and supervised finetuning setups.

Further, we provide additional implementation details on our evaluation setup of video backbones in action classification and text-to-video retrieval. To evaluate a video backbone in action classification, following the VideoGLUE benchmark [38], we train a classification model with the corresponding frozen video backbone and single-layer pooler head [38] with one classification linear layer as described in the main paper. For training and evaluation, we use 8 uniformly sample frames as inputs, however, for video backbones that create 3D tokens with a stride over frames, such as VideoMAE [32, 34], we respectively scale the number of input frames, namely, we use 16 frames for VideoMAE and VideoMAEv2, to ensure that models use the same number of tokens for the same model size. We train a model for 50 epochs with an AdamW optimizer [22] and a learning rate of 0.001. We use a cosine weight decay scheduler with five epochs warmup. We follow the augmentation pipeline of the VideoGLUE benchmark [38]. We do not use multi-crop evaluation to simplify and standardize the evaluation setting for all datasets.

For text-to-video retrieval, we evaluate the zero-shot capabilities of the text-video models and use the respective backbones without fine-tuning. Same as in action classification, we use 8 uniformly sampled frames as inputs. We also follow the corresponding model recipes in using reranking, namely, we rerank 128 videos with highest similarities based on dual encoder output by a joint encoder for Unmasked Teacher (UMT) [16] and VideoMamba [17] models.

C. Qualitative Results

C.1. UTD-descriptions

In Figs. F.1 to F.4, we present qualitative results of textual descriptions of our *UTD-descriptions* dataset for different concept categories using random videos from the MSRVTT dataset. We observed that the VLM provides detailed descriptions of objects+composition+activities (Fig. F.1). Furthermore, the LLM successfully parses these descriptions

Model	Backbone	Pretraining Datasets	Finetuning Datasets
VideoMAE-B-K400 [32]	ViT-B/16	Kinetics-400 (w/o labels) [13]	-
VideoMAE-B-UH [32]	ViT-B/16	UnlabeledHybrid [34]: K700 [6] + WebVid2M [2] + SS [10] + AVA [11] + Instagram (collected)	-
VideoMAE-L-UH [32]	ViT-L/14	UnlabeledHybrid [34]: K700 [6] + WebVid2M [2] + SS [10] + AVA [11] + Instagram (collected)	-
VideoMAE-H-UH [32]	ViT-H/16	UnlabeledHybrid [34]: K700 [6] + WebVid2M [2] + SS [10] + AVA [11] + Instagram (collected)	-
VideoMAEv2-B-K710-fnK710 [34]	ViT-B/16	Kinetics-710 [15] (Kinetics-400 [13] + Kinetics-600 [5] + Kinetics-700 [6]) (w/o labels)	Kinetics-710 [15]
AllInOne-B-WV2M+CC [33]	ViT-B/16	WebVid2M [2] + CC3M [30]	-
AllInOne-B-WV2M+HT [33]	ViT-B/16	WebVid2M [2] + HowTo100M [24]	-
AllInOne-B-WV2M+HT+CC+YTT+ [33]	ViT-B/16	WebVid2M [2] + HowTo100M [24] + YTT [39] + CC3M [30] + CC12M [7] + COCO [18] + VG [14] + SBU [27]	-
UMT-B-K710 [16]	UMT-B/16	Kinetics-710 [15] (w/o labels)	-
UMT-B-fnK710 [16]	UMT-B/16	Kinetics-710 [15] (w/o labels)	Kinetics-710 [15]
UMT-L-K710 [16]	UMT-L/16	Kinetics-710 [15] (w/o labels)	-
UMT-L-fnK710 [16]	UMT-L/16	Kinetics-710 [15] (w/o labels)	Kinetics-710 [15]
UMT-B-5M [16]	UMT-B/16	Kinetics-710 [15] (w/o labels) + WebVid2M [2] + CC3M [30]	-
UMT-B-17M [16]	UMT-B/16	Kinetics-710 [15] (w/o labels) + WebVid2M [2] + CC3M [30] + CC12M [7] + COCO [18] + VG [14] + SBU [27]	-
UMT-B-25M [16]	UMT-B/16	Kinetics-710 [15] (w/o labels) + WebVid10M [2] + CC3M [30] + CC12M [7] + COCO [18] + VG [14] + SBU [27]	-
UMT-L-5M [16]	UMT-L/16	Kinetics-710 [15] (w/o labels) + WebVid2M [2] + CC3M [30]	-
UMT-L-17M [16]	UMT-L/16	Kinetics-710 [15] (w/o labels) + WebVid2M [2] + CC3M [30] + CC12M [7] + COCO [18] + VG [14] + SBU [27]	-
UMT-L-25M [16]	UMT-L/16	Kinetics-710 [15] (w/o labels) + WebVid10M [2] + CC3M [30] + CC12M [7] + COCO [18] + VG [14] + SBU [27]	-
VideoMamba-VM-K400 [17]	VideoMamba-M	Kinetics-400 [13] (w/o labels)	-
VideoMamba-VM-5M [17]	VideoMamba-M	Kinetics-400 [13] (w/o labels) + WebVid2M [2] + CC3M [30]	-
VideoMamba-VM-17M [17]	VideoMamba-M	Kinetics-400 [13] (w/o labels) + WebVid2M [2] + CC3M [30] + CC12M [7] + COCO [18] + VG [14] + SBU [27]	-
VideoMamba-VM-25M [17]	VideoMamba-M	Kinetics-400 [13] (w/o labels) + WebVid10M [2] + CC3M [30] + CC12M [7] + COCO [18] + VG [14] + SBU [27]	-
InternVid-B-10M-FLT [36]	ViCLIP-B/16	InternVid-10M-FLT [36]	-
InternVid-B-200M [36]	ViCLIP-B/16	InternVid-200M [36]	-
InternVid-L-10M [36]	ViCLIP-L/14	InternVid-10M [36]	-
InternVid-L-WebVid10M [36]	ViCLIP-L/14	WebVid10M [2]	-
InternVid-L-10M-DIV [36]	ViCLIP-L/14	InternVid-10M-DIV [36]	-
InternVid-L-10M-FLT [36]	ViCLIP-L/14	InternVid-10M-FLT [36]	-
InternVid-L-50M [36]	ViCLIP-L/14	InternVid-50M [36]	-
InternVid-L-200M [36]	ViCLIP-L/14	InternVid-200M [36]	-

Table B.1. Overview of the 30 models analyzed in this paper, including their backbones, pretraining datasets, and supervised finetuning datasets.

into objects, activities, and verbs (Figs. F.2 to F.4).

C.2. UTD-splits

In Fig. F.5, we present qualitative examples from the full test set and our object-debiased UTD-split on the UCF101 dataset. We observe that samples from our object-debiased UTD-split demand a deeper level of video understanding beyond simple object recognition. For instance, in examples of classes involving playing musical instruments, such as "Playing Daf", "Playing Cello", or "Playing Sitar", the videos often include additional instruments in the background, such as a piano and drums in the case of "Playing Daf" example, alongside the primary instrument. Similarly, in the "Pizza Tossing" example, the pizza in the UTD-split example is barely visible, and a video requires analysis beyond this single frame for correct class prediction.

C.3. User Study

As stated in the main paper, our user study shows that 87.6% (667 out of 761) of the VLM-recognized objects are identified as visible, with only 94 objects not selected as visible. To better understand the VLM's errors, we manually classifiy these 94 objects into five categories: 1) attribute error (e.g., the object is "right hand" instead of "left hand"), 2) misclassification (the object is present but incorrectly identified as a different object, e.g., a "snowboard" instead of "snow slide"), 3) hallucination, 4) human annotation mistake – the object is visible, and 5) other. In Fig. F.6 we provide examples from these five categories.

D. License Information

Our *UTD-descriptions* are generated by LLaVA-1.6-7B-mistral model [19, 21] as well as Mistral-7B-Instruct-v0.2 [12]. The models were used according to their licenses. LLaVA-1.6-7B-mistral model complies with the base LLM's model license[1], which is Mistral-7B. Mistral-7B model follows Apache 2.0 license[2]. We release our *UTD dataset*, comprising *UTD-descriptions* and *UTD-splits*, under the CC BY 4.0 license. However, specific components of the underlying dataset may be governed by stricter licensing conditions from the corresponding video datasets.

E. Limitations and Broader Impact

Limitations. Due to the high human annotation cost of detailed frame text descriptions for videos, our method of evaluating and debiasing datasets relies on the textual description of video frames generated by VLMs. Therefore, these textual descriptions may potentially misrepresent the video due to the potential limitation of VLM models. Namely, such textual description might be prone to hallucinations, namely describing things that are not present in the frame, having social biases, such as guessing a person's occupation based on how the person looks or implying some information that is not visible. Then, we further extract different concept categories from these textual descriptions us-

[1] https://github.com/haotian-liu/LLaVA/blob/main/docs/MODEL_ZOO.md

[2] https://mistral.ai/news/announcing-mistral-7b/

ing LLM. These steps might be prone to leaking information about other concepts, such as adding activity information to the objects list. Therefore, the quality of our textual descriptions with respect to different concepts also limited the filtering abilities of LLMs. In our user study evaluating the quality of generated descriptions of objects, we found the hallucination rate to be approximately 6%. Our biased discovering method is also limited by the performance of pretrained text embedding models, and using stronger text embedding models might lead to discovering even stronger biases in the datasets, such as more videos being classified correctly based on only objects using common sense.

Potential Negative Societal Impact. In our work, we perform bias analysis of the existing video classification and text-to-video retrieval datasets, 12 datasets in total, as well as provide debiased evaluation splits – *UTD-splits*. Since our annotations are based on existing datasets, our data distribution reflects the social biases inherent in those sources. We also provide *UTD-descriptions* for these 12 datasets – textual descriptions of objects, activities, verbs, and objects+composition+activities categories. As discussed in the limitations, these descriptions might be prone to social biases potentially present in the current VLM models. Therefore, such descriptions might potentially propagate these social biases. To create the dataset, we utilize VLM and LLM models, which contribute to increased energy consumption and carbon emissions as a negative externality.

F. UTD Dataset

We release the UTD dataset, which consists of two parts:

1. UTD**-descriptions.** This includes frame annotations for four conceptual categories visible in video frames: *objects, activities, verbs*, and *objects+composition+activities*. UTD-descriptions are provided for 8 uniformly sampled frames from the training and test/val sets of 12 action recognition and text-to-video retrieval datasets. The annotations *for objects+composition+activities* are generated using the LLaVA-1.6-7B-mistral VLM prompted to describe visible object relationships in a frame. From these descriptions, *objects, activities*, and *verbs (activities without associated objects)* are derived using the Mistral-7B-Instruct-v0.2 model.
2. UTD**-splits.** This includes object-debiased test/val splits, which are subsets of the original test/val splits with object-biased items removed. These debiased splits are provided for the 12 considered activity recognition and text-to-video retrieval datasets. For the 6 activity recognition datasets, we additionally provide debiased-balanced splits, where the most object-biased samples are removed while preserving the original class distribution to ensure fair evaluation across categories.

The download instructions, documentation, and usage guidance may be found on our project webpage: `https://utd-project.github.io/` Below, we provide the datasheet for our UTD dataset, license information, and statement of responsibility.

UTD Dataset Datasheet

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The UTD Dataset is a benchmark designed to assess the performance of video backbones. It consists of debiased evaluation subsets, specifically video IDs, of 12 popular action classification and text-to-video retrieval datasets (UTD-splits), namely UCF101 [31], SomethingSomethingv2 [10], Kinetics-400 [13], -600 [5], and -700 [6], and Moments In Time [25], MSRVTT [37], DiDeMo [1], ActivityNet [4], LSMDC [29], YouCook2 [8], and Spoken Moments In Time [26]. The goal is to evaluate the robustness of video models to object representation bias and to provide a challenging benchmark for evaluating video models with reduced object bias in the evaluation set. While previous work has focused on assessing and mitigating various representation biases in video benchmarks, debiased solutions have rarely been adopted for benchmarking. This is due to several reasons, such as the additional training and/or testing overhead required or the necessity to address out-of-domain problems. Our work introduces a novel method for evaluating and debiasing existing datasets via their textual descriptions. This approach allows us to identify and remove samples with object representation bias from the evaluation sets. Additionally, the dataset includes UTD-descriptions, which are textual descriptions of four conceptual categories visible in video frames: objects, activities, verbs, and objects+composition+activities. These annotations cover the 12 corresponding datasets and aim further to advance the measurement of representation biases in the field.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by a research group affiliated with the Goethe University Frankfurt, Tuebingen AI Center/University of Tuebingen, University of Oxford, MPI for Informatics, and MIT-IBM Watson AI Lab.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Individual researchers within the research group have been funded by the German Federal Ministry of Education and Research (BMBF) project STCL - 01IS22067, the ERC Starting Grant GraViLa 101117556, and supported by travel grants from ELISE (GA no 951847).

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Our dataset builds upon existing datasets. It contains only textual annotation or meta-annotation. UTD-splits contain lists of video IDs. UTD-descriptions contain texts.

How many instances are there in total (of each type, if appropriate)?

UTD-splits contain debiased splits for 12 different action classifications and text-to-video retrieval datasets. UTD-descriptions contain textual annotation for train/test videos of corresponding datasets, describing $\sim$1.9M videos in total.

Does the dataset contain all possible instances, or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances because instances were withheld or unavailable).

Our dataset builds upon existing datasets. It provides video IDs of the samples that are more representative of the corresponding tasks and samples that cannot be easily solved with simple techniques. The main purpose of this dataset is to filter out these non-representative, easy samples from existing datasets.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

An instance of UTD-splits is the name of a video dataset along with a list of video IDs in the object-debiased subsets and, for six activity recognition datasets, a list of video IDs in the object-debiased-balanced subsets.

An instance of UTD-descriptions is a video ID, for which we provide annotations for four conceptual categories: objects, activities, verbs, and objects+composition+activities. These annotations are provided for 8 uniformly sampled frames for video corresponding to the video ID.

Is there a label or target associated with each instance? If so, please provide a description.

Not applicable. We provide annotations for existing datasets that already have established labels. These type of labels varies across datasets and tasks.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Instances are complete.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Not applicable.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Yes. We provide annotations for existing datasets with well-established splits.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

UTD-descriptions are generated by VLM and LLM models, which might introduce hallucinations, social biases, or imply information that is not actually visible in the frames. Since UTD-splits are derived using UTD-descriptions, it too may be susceptible to these errors.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is publicly available on our project webpage: https://utd-project.github.io/. This webpage includes links to data files hosted on Google Drive, and long-term public accessibility and maintenance will be ensured. The dataset will be released under the CC-4.0 license. However, certain parts of the upstream datasets may be subject to stricter licensing conditions from the corresponding video datasets.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

There is a small chance that the automatically generated text annotations can contain offensive language. However, with extensive manual checks, we have not encountered such a sample.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Not applicable.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Since UTD-splits contain only video IDs and UTD-descriptions provide textual descriptions of video frames, there is a very low chance that PID will be captured in the annotations.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

UTD-descriptions are generated by VLM and LLM models from video frames. UTD-splits are derived using UTD-descriptions.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

Our dataset builds upon 12 existing datasets, namely UCF101 [31], SomethingSomethingv2 [10], Kinetics-400 [13], -600 [5], and -700 [6], and Moments In Time [25], MSRVTT [37], DiDeMo [1], ActivityNet [4], LSMDC [29], YouCook2 [8], and Spoken Moments In Time [26]. To compile our dataset, we first downloaded videos from these 12 datasets following their official instructions. We then generated UTD-descriptions using the officially released LLaVA-1.6-7B-mistral model and Mistral-7B-Instruct-v0.2 on an internal cluster. Additionally, we derived UTD-splits using the officially released SFR-Embedding-Mistral model. The detailed methodology is provided in the paper.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Not applicable.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Not applicable.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The dataset was created in 2024.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

We did not perform any new data collection process but utilized data from 12 existing datasets, using corresponding official instructions to access the data.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Our dataset is a meta-dataset and thus, by itself, does not collect any new data. All 12 considered datasets are publicly available and contain videos sourced from publicly accessible resources such as YouTube and other internet platforms, consisting of user uploads, however, we are not aware whether consent was obtained from the users.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Please see the previous answer.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Not applicable.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

We provided a simple post-processing of objects, activities, and verbs textual descriptions, such as removing numeration and text in brackets.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Yes, we provide a link to a raw version in the project webpage.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

We used only simple Python scripts for this which we release.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

In our paper, we demonstrate the intended use of UTD-descriptions by deriving UTD-splits. We also use UTD-splits to benchmark various video backbones and analyze their robustness to object bias. All details can be found in the paper.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

All links are provided in the paper.

What (other) tasks could the dataset be used for?

We believe that UTD-descriptions, which contain dense textual descriptions (for 8 uniformly sampled frames) of different concept categories—namely objects, activities, verbs, and objects+composition+activities—for 12 popular video datasets, could be widely used by the community for various tasks. Examples of other uses include deriving new datasets or models for understanding object relationships in videos or creating new challenging VQA datasets that require temporal understanding.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

Our dataset provides annotations for existing datasets and is intended to be used in conjunction with those datasets. Therefore, while using videos and other data from the original datasets, users should comply with the licenses and terms of usage of these datasets, which are mostly restricted to research purposes. Additionally, since our annotations are generated using models, users should be aware of potential biases and inaccuracies and take appropriate measures to mitigate any risks or harms.

Are there tasks for which the dataset should not be used? If so, please provide a description.

The dataset should be used for research only.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

No.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be provided on our project webpage: `https://utd-project.github.io/`. This repository includes links to JSON data files hosted on Google Drive. It does not currently have a DOI.

When will the dataset be distributed?

The dataset will be distributed starting in March 2025.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset will be released under the Creative Commons Attribution 4.0 (CC BY 4.0) license. The terms of this license can be found at: `http://creativecommons.org/licenses/by/4.0`. However, certain parts of the underlying dataset may be subject to stricter licensing conditions from the corresponding video datasets.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Not that we are aware of.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset will be supported and maintained by the authors of the paper. The main contact person is Nina Shvetsova.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The authors can be contacted via the following email addresses: {*shvetsov,kuehne*}*@uni-frankfurt.de*, *arsha@robots.ox.ac.uk*, *schiele@mpi-inf.mpg.de*, *christian.rupprecht@cs.ox.ac.uk*.

Is there an erratum? If so, please provide a link or other access point.

Errata will be posted on the project's webpage.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

Updates will be communicated through the project's webpage and will be versioned. We will strive to correct errors promptly and may add or delete instances as necessary.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Yes, we will delete instances upon request.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

The older versions of the dataset will continue to be hosted on Google Drive. They will remain accessible through the project's webpage where updates and newer versions will also be posted.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

We welcome contributions and ideas from others who wish to extend, augment, or build upon our dataset. Interested parties can reach out to us via email to discuss their ideas.

	Prompt
Prompting LLaVA-1.6-Mistral-7B: Obtaining Objects+Composition+Activities $d_{n,i}$	`Describe the objects relationships in the photo.`
Prompting Mistral-7B-Instruct-v0.2: Obtaining Objects $o_{n,i}$	`<s>[INST] You are an intelligent chatbot designed to extract requested information from the textual description of an image . I will give you a textual description of the image. List ALL objects visible in the image. An object is anything that has a fixed shape or form, that you can touch or see. Name each object with one noun or a maximum of two words. Skip uncertain objects. The textual description of the image: "<INPUT TEXTUAL DESCRIPTION>" DO NOT PROVIDE ANY EXTRA INFORMATION ABOUT OBJECT PROPERTIES OR RELATIONSHIPS TO OTHER OBJECTS IN PARENTHESES. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. [/INST] Comprehensive enumerated list of objects:`
Prompting Mistral-7B-Instruct-v0.2: Obtaining Activities $a_{n,i}$	`<s>[INST] You are an intelligent chatbot designed to extract requested information from the textual description of an image . I will give you a textual description of the image. List all VISIBLE activities in the image. Activity is lively action or movement. Name each activity with a concise phrase SKIP possible or implied activities that are not visible. If no activity is visible, reply "No activity is visible." DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. The textual description of the image: "<INPUT TEXTUAL DESCRIPTION>" [/INST] Comprehensive enumerated list of activities:`
Prompting Mistral-7B-Instruct-v0.2: Obtaining Verbs $\nu_{n,i}$	`<s>[INST] You are an intelligent chatbot designed to extract requested information from the textual description of an image . I will give you a list of visible activities of the image. Your task is to delete information about objects from this description. Replace all objects in this list with "someone" or "something," but keep the activity. If you have to, you may delete some details, but delete ALL object information. If the input is "No activity is visible.", keep it "No activity is visible." DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. The list of visible activities: "<INPUT ACTIVITIES DESCRIPTION >" [/INST] Post-processed enumerated list of activities:`
Prompting Mistral-7B-Instruct-v0.2: Obtaining 15-words Summaries $d'_{n,i}$	`<s>[INST] You are an intelligent chatbot designed to extract requested information from the textual description of an image. Summarize the following image description in 15 words: "<INPUT TEXTUAL DESCRIPTION>" [/INST] 15-words summary:`

Table F.1. Prompts used in our UTD method to obtain textual descriptions of frames with respect to different concepts categories: objects+composition+activities $d_{n,i}$, objects $o_{n,i}$, activities $a_{n,i}$, and verbs $\nu_{n,i}$.

Textual description	Setup	<instruction>
objects+composition+activities $d_{n,i}$	single-frame	Given a video frame description, retrieve the activity depicted in this video.
	sequence-of-frames	Given descriptions of video frames, retrieve the activity depicted in this video.
objects $o_{n,i}$	single-frame	Given a list of objects visible on the video frame, retrieve the activity depicted in this video.
	sequence-of-frames	Given lists of objects visible on the video frames, retrieve the activity depicted in this video.
activities $a_{n,i}$ / verbs $\nu_{n,i}$	single-frame	Given a description of actions visible on the video frame , retrieve the activity depicted in this video.
	sequence-of-frames	Given a description of actions visible on the video frames, retrieve the activity depicted in this video.
activity class name	single-frame	Given an activity, retrieve a video frame description that may depict this activity.
	sequence-of-frames	Given an activity, retrieve a video description that may depict this activity.

Table F.2. Instructions used to prompt the SFR-Embedding-Mistral model for action classification.

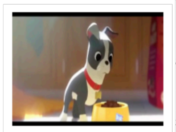
Textual description	Setup	<instruction>
objects+composition+activities $d_{n,i}$	single-frame	Given a description of a single video frame, retrieve a short description of the full video.
	sequence-of-frames	Given descriptions of video frames, retrieve a short description of the full video.
objects $o_{n,i}$	single-frame	Given a list of objects visible on the video frame, retrieve a short video description.
	sequence-of-frames	Given lists of objects visible on the video frames, retrieve a short video description.
activities $a_{n,i}$ / verbs $\nu_{n,i}$	single-frame	Given a description of actions visible on the video frame , retrieve a short video description.
	sequence-of-frames	Given a description of actions visible on the video frames, retrieve a short video description.
caption (from dataset)	single-frame	Given a short video description, retrieve a description of a specific frame within that video.
	sequence-of-frames	Given a short video description, retrieve another description of this video.

Table F.3. Instructions used to prompt the SFR-Embedding-Mistral model for text-to-video retrieval.

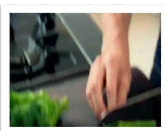
Task	Textual Description	`<instruction>`
retrieval	objects o_n (seq-of-frames setup)	`Given lists of objects visible on the video frames, retrieve a short video description.`
		`Using lists of objects seen in video frames, retrieve a brief description of the video.`
		`From lists of objects present in video frames, retrieve a concise video description.`
	caption (from dataset)	`Given a short video description, retrieve another description of this video.`
		`Use a brief video description as a query to retrieve an alternative description of the same video.`
		`Given a concise video description, retrieve another description for that video.`
classification	objects o_n (seq-of-frames setup)	`Given lists of objects visible on the video frames, retrieve the activity depicted in this video.`
		`Using lists of objects seen in video frames, retrieve the activity captured in the video.`
		`From lists of objects present in video frames, retrieve the activity that the video shows.`

Table F.4. Multiple instructions for robust datasets unbiasing. We prompt the text embedding models using three different prompts for captions and object textual descriptions, generating three different embeddings for each. During the unbiasing process, we exclude samples from the test sets only if the sample is correctly classified or retrieved using any combination of these embeddings.

Objects+composition+activities:
In the image, there is a person standing on the bow of a small boat that is floating on a body of water. The person appears to be engaged in an activity that involves a blue, translucent object that is being pulled or dragged by the boat. The object is connected to the boat by a line, and the person seems to be controlling the movement of the boat, which in turn affects the movement of the object. The background shows a calm water surface with trees and a clear sky, suggesting a peaceful, outdoor setting. The text overlay on the image provides information about the person and the activity, indicating that the person is involved in "Authentic Low Country Cooking" and that the image is from a website called "gullahgrubs.com." The overall scene suggests a leisurely activity related to cooking, possibly involving the use of the blue object for food preparation or demonstration.

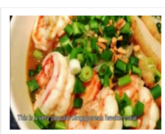


Objects+composition+activities:
In the image, there is an animated dog character standing next to a yellow bucket. The dog appears to be looking at the bucket, which has a label on it. The dog is wearing a red collar with a tag, and it has a black and white coat with a white muzzle. The background suggests an indoor setting with a wooden floor and a blurred object that could be a bottle or a container in the background. The dog's position relative to the bucket is such that it seems to be interacting with the contents of the bucket, possibly trying to get something out of it. The label on the bucket is not clearly legible, so its contents are not identifiable from the image.

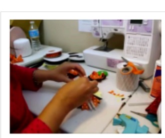


Objects+composition+activities:
In the photo, there is a person's hand holding a black object with a green label. The object appears to be a bottle or container of some sort. The hand is positioned in such a way that it is holding the object securely. The background is blurred, but it seems to be a kitchen setting, as suggested by the presence of what looks like a kitchen counter or table. The focus of the image is on the hand and the object it is holding, with the background being out of focus. The relationship between the objects is that the hand is interacting with the object, possibly opening it or preparing to use it.

Objects+composition+activities:
The image contains two lines of text. The first line reads "An easy folding stroller" and the second line reads "will make." The text is presented in a simple, sans-serif font, and the color of the text is a shade of blue. The relationship between the two lines of text is that the first line is a statement, and the second line is a continuation of the thought or instruction implied by the first line. The second line is a verb, suggesting an action that will be taken or the result of using the stroller mentioned in the first line. The text is centered on the page, and there are no other objects or relationships to describe.



Objects+composition+activities:
The image shows a bowl of food that appears to be a type of noodle dish, possibly a Singaporean hawker meal. The dish is garnished with various ingredients, including slices of what could be a type of seafood, such as shrimp or prawns, green onions, and what looks like fried shallots or garlic. The noodles are likely to be a type of egg noodle, given their appearance and the context of the dish.
The ingredients are arranged in a way that suggests a balance of flavors and textures. The seafood provides a protein source and a hint of the sea, while the green onions add a fresh, herbal flavor. The fried shallots or garlic contribute a savory, slightly sweet element to the dish.
The dish is presented in a white bowl, which contrasts with the colorful ingredients, making the food look more appetizing. The bowl is placed on a surface that is not clearly visible, but it appears to be a table or countertop.
The text overlay on the image provides a description of the dish, indicating that it is a popular Singaporean hawker meal. This suggests that the image is likely intended for a culinary context, such as a food blog or a recipe sharing platform. The text is in English and is superimposed over the image, ensuring that the viewer can read it regardless of the image's resolution or size.



Objects+composition+activities:
In the photo, there is a person sitting at a table with various crafting supplies. The person appears to be engaged in a crafting activity, possibly sewing or working with fabric. The table is cluttered with several items, including:
1. A sewing machine, which is likely being used by the person.
2. A spool of thread, which is being used by the person.
3. A pair of scissors, which are commonly used in sewing and crafting.
4. A roll of tape, which could be used for measuring or securing items.
5. A pair of pliers, which might be used for cutting or bending materials.
6. A bottle of water, which is likely there for the person to drink.
7. A roll of paper towels, which could be used for cleaning up any messes.
8. A pair of tweezers, which might be used for picking up small items or threads.
9. A pair of scissors with orange handles, which could be a different type of scissors or a different color for a specific purpose.
10. A small blue object, which could be a tool or a small part of the crafting project.

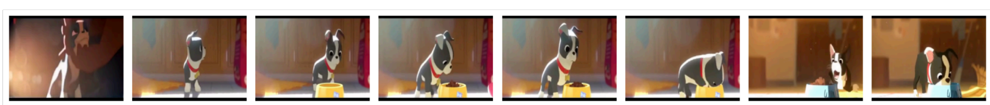
The objects are arranged in a way that suggests an active workspace, with the person using the sewing machine and the other items within easy reach. The items are not neatly organized, indicating that the person is in the middle of a project and has not yet tidied up the workspace.

Figure F.1. Qualitative examples of objects+composition+activities textual descriptions for random videos in MSRVTT dataset.

Objects:
Frame 1: person, small boat, body of water, blue object, string or line
Frame 2: person, small boat, bow, blue, translucent object, string, wide-brimmed hat, patterned shirt, body of water, trees, blue smoke or vapor, text overlay "gullah grub", "authentic low country cooking", "www
Frame 3: person, hat, small boat, bow, body of water, blue object, trees, clear sky.
Frame 4: person, small boat, bow, body of water, blue, translucent object, trees, clear sky.
Frame 5: person, small boat, body of water, blue, translucent object, line, trees, clear sky, text overlay, website name "gullahgrubs.com"
Frame 6: person, hat, small boat, body of water, bird, string or line
Frame 7: person, small boat, bow, fishing rod, line, body of water, trees, clear sky, text: "produced & directed by:", text: "www.outsidetheboxmedia.com", text: "copyright 2011"
Frame 8: person, hat, small boat, body of water, bird, string or line.

Activities:
Frame 1: A person is holding a blue object. The person is pulling on a string or line attached to the blue object.
Frame 2: A person is controlling the movement of a blue, translucent object using a string.
Frame 3: A person is standing on a boat. A person is pulling a blue object through water.
Frame 4: A person is standing on the bow of a boat. A person is spinning or twirling a blue, translucent object.
Frame 5: A person is controlling a boat. An object is being pulled or dragged by the boat.
Frame 6: A person is controlling a bird's flight with a string or line.
Frame 7: A person is standing on the bow of a boat. A person is fishing or participating in a water-based activity using a fishing rod.
Frame 8: A person is controlling a bird's flight with a string or line.

Verbs:
Frame 1: Someone is holding something. Someone is pulling on something.
Frame 2: Someone is moving something using something.
Frame 3: Someone is standing on something. Someone is pulling something through water.
Frame 4: Someone is standing. Someone is spinning or twirling something.
Frame 5: Someone is controlling something. Something is being pulled or dragged.
Frame 6: Someone is controlling something.
Frame 7: Someone is standing. Someone is participating in an activity using a rod.
Frame 8: Someone is controlling something.

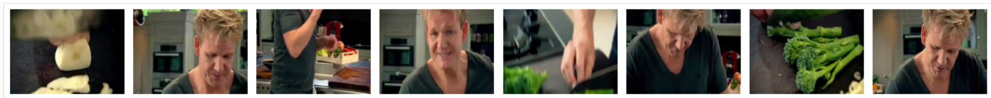


Objects:
Frame 1: character, red and white striped shirt, room, warm, golden light, fireplace or glowing light source
Frame 2: animated character, dog, wooden door, red object, bottle or container, label, bell
Frame 3: animated dog, yellow bucket, wooden floor, red collar, red object
Frame 4: 3d animated dog, yellow bucket, surface, blurred background, wooden door, red object
Frame 5: animated dog, yellow bucket, red collar, tag, wooden floor, blurred object
Frame 6: cartoon dog, yellow object, container or box, indoor setting, wooden surface, red object, bottle or container.
Frame 7: cartoon dog, blue bowl, wooden door, window, curtains.
Frame 8: dog, blue bowl, wooden floor, furniture

Activities:
Frame 1:
Frame 2: The dog is looking towards the door. The dog might be waiting or anticipating something.
Frame 3: The dog is looking at the bucket.
Frame 4: The dog is interacting with the bucket. The dog is possibly sniffing or looking inside the bucket.
Frame 5: The dog is looking at the bucket. The dog is interacting with the contents of the bucket.
Frame 6: The dog is looking towards the right side. It is unclear if the dog is interacting with the yellow object or just standing next to it. Therefore, the activity is listed as "Interacting with the yellow object " to account for both possibilities.
Frame 7: The dog is standing on its hind legs. The dog is looking at the blue bowl. The dog might be trying to reach the contents of the bowl.
Frame 8: The dog is leaning over the bowl. The dog is eating or drinking from the bowl.

Verbs:
Frame 1:
Frame 2: Something is looking. Something might be waiting or anticipating.
Frame 3: Something is looking at something.
Frame 4: Something is interacting with something. Something is investigating something.
Frame 5: Something is looking at something. Something is interacting with something.
Frame 6: Something is looking towards something. Someone is interacting with something .
Frame 7: Something is standing on its hind legs. Something is looking at something. Something might be trying to reach something.
Frame 8: Something is leaning. Something is consuming.

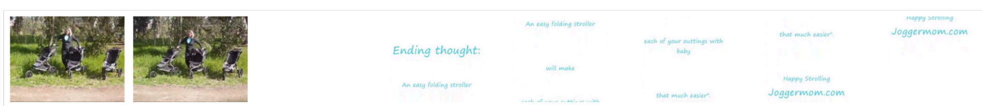
Figure F.2. Qualitative examples of objects, activities, and verbs textual descriptions for random videos in MSRVTT dataset.



Objects:
Frame 1: hand, food, potato or similar starchy vegetable, peel, background
Frame 2: man, dark-colored top, kitchen counter, bottle, bowl, large appliance, refrigerator or freezer, handle, kitchen cabinets, artificial lighting.
Frame 3: person, dark green t-shirt, cutting board, vegetables, red toaster, white container, countertop, bowl, spoon.
Frame 4: man, green top, black oven, glass door, white cabinet, handle, black refrigerator, freezer compartment, shelf, cans.
Frame 5: hand, black object, green label, kitchen counter or table.
Frame 6: man, dark green/black t-shirt, red appliance, window, garden or outdoor area, countertop, bottle, bowl.
Frame 7: cutting board, broccoli, knife, countertop.
Frame 8: person, man, dark-colored shirt, kitchen counter, kitchen cabinet, microwave, kitchen appliance.

Activities:
Frame 1: A person is peeling a potato or a similar starchy vegetable.
Frame 2:
Frame 3: The person is cooking or preparing food.
Frame 4:
Frame 5: A hand is holding an object. The hand is interacting with the object.
Frame 6: The man is looking down.
Frame 7: Cutting broccoli with a knife.
Frame 8: The man is looking at a piece of paper or a recipe. The man is standing in front of a kitchen appliance, suggesting he might be preparing to use it.

Verbs:
Frame 1: Someone is peeling something.
Frame 2:
Frame 3: Someone is cooking or preparing.
Frame 4:
Frame 5: Something is holding something. Something is interacting with something.
Frame 6: Someone is looking down.
Frame 7: Someone is cutting something.
Frame 8: Someone is looking at something. Someone is preparing to use something.



Objects:
Frame 1: woman, stroller, stroller, black frame, canopy, grassy area, trees.
Frame 2: woman, stroller, stroller, grassy area, trees.
Frame 3: white background, darker line.
Frame 4: text, subtitle, main text, font, white background.
Frame 5: text, two lines, sans-serif font, blue text
Frame 6: text message, light blue background, darker blue font, text: "each of your outtings with baby that much easier.", text: "that much easier."
Frame 7: text line 1, text line 2, sans-serif font, shade of blue
Frame 8: text, "happy", "strolling", "joggermom", ".", "com"

Activities:
Frame 1: The woman is holding onto the handles of both strollers. The woman is either pushing or maneuvering the strollers.
Frame 2: The woman is maneuvering or pushing the strollers.
Frame 3:
Frame 4:
Frame 5:
Frame 6:
Frame 7:
Frame 8:

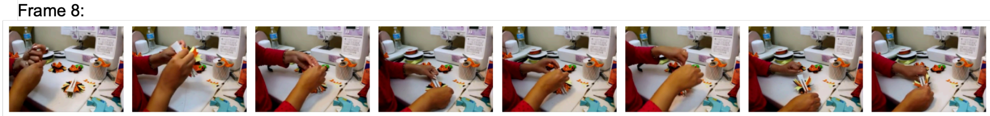
Verbs:
Frame 1: Someone is holding onto something. Someone is maneuvering something.
Frame 2: Someone is maneuvering or pushing.
Frame 3:
Frame 4:
Frame 5:
Frame 6:
Frame 7:
Frame 8:

Figure F.3. Qualitative examples of objects, activities, and verbs textual descriptions for random videos in MSRVTT dataset.

Objects:
Frame 1: bowl, noodles, prawns, green onions, red chili peppers, fried shallots or garlic, sauce, spoon.
Frame 2: bowl, singaporean prawn noodle soup, green onions, red chili peppers, fried shallots or garlic, prawns, noodles, light-colored liquid, text overlay "singapore prawn noodle soup", surface
Frame 3: bowl, seafood, sauce, green onions, chopped nuts.
Frame 4: bowl, noodles, shrimp, green onions, green bell pepper, chopped green onions
Frame 5: bowl, noodle dish, seafood, shrimp or prawns, green onions, fried shallots or garlic.
Frame 6: bowl, seafood dish, green onions, red chili peppers, fried shallots or garlic.
Frame 7: seafood stir-fry, shrimp, green onions, chopped nuts, rice.
Frame 8: plate, seafood dish, shrimp, green onions, chopped green onions, fried shallots or garlic, white plate, text overlay, rich broth.

Activities:
Frame 1:
Frame 2:
Frame 3:
Frame 4:
Frame 5:
Frame 6:
Frame 7:
Frame 8:

Verbs:
Frame 1:
Frame 2:
Frame 3:
Frame 4:
Frame 5:
Frame 6:
Frame 7:
Frame 8:



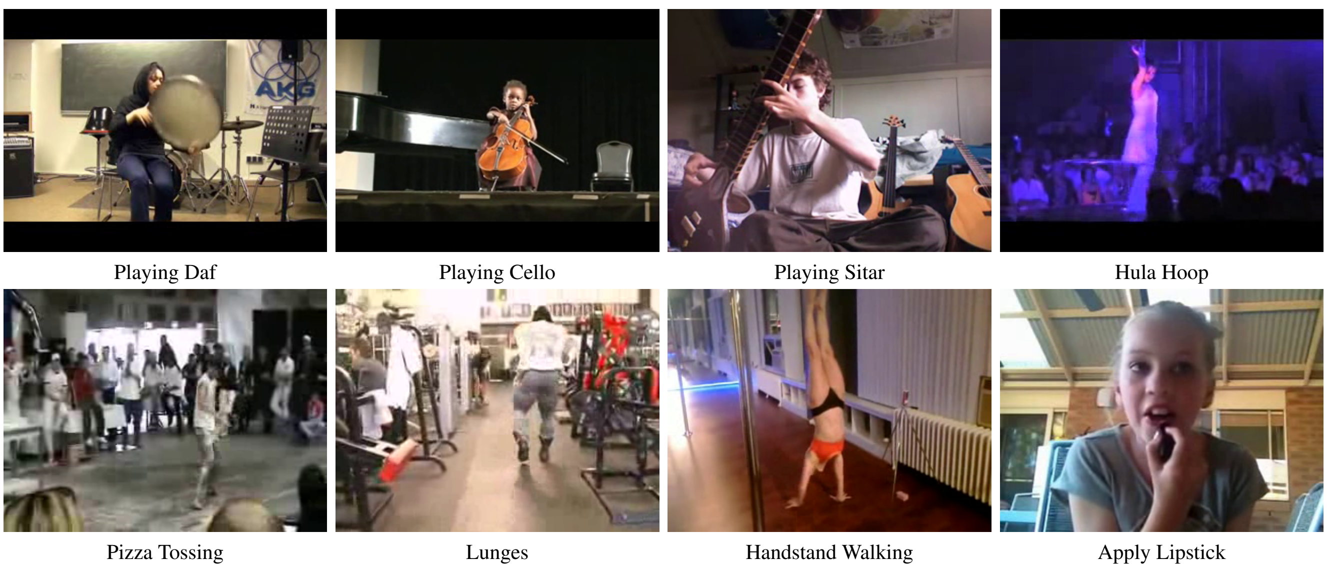
Objects:
Frame 1: person, table, sewing machine, pair of scissors, roll of tape, pair of tweezers, bottle of water, roll of paper towels, pair of pliers, pair of earrings.
Frame 2: person, table, sewing machine, roll of thread or yarn, pair of scissors, pliers, piece of fabric or paper with a design, bottle of water, roll of tape.
Frame 3: person, sewing machine, red top, fabric, thread, table or countertop, white plastic container, roll of thread, pair of scissors, bottle of water, roll of fabric or paper.
Frame 4: person, sewing machine, bottle of water, roll of fabric, pair of scissors, spool of thread, measuring tape, white printer, table
Frame 5: person, table, sewing machine, spool of thread, pliers, bottle of water, roll of paper towels, pair of scissors, small blue object, pair of scissors, spool, pair
Frame 6: person, sewing machine, fabric or paper scraps, small object, table, pair of scissors, orange handles, roll of clear tape, bottle of water, blue lid, pliers, small blue object.
…

Activities:
Frame 1: The person is crafting. The person is using a sewing machine. The person is using scissors. The person is using tweezers. The person is using pliers. The person is using glue. The person is possibly cleaning up with paper towels. The person is possibly drinking water.
Frame 2: A person is sitting and working. A person is using a sewing machine. A person is cutting with scissors.
Frame 3: A person is sitting and using a sewing machine. A person is working on a sewing project. A person is storing scissors in a container. A person is possibly cutting materials with scissors .
Frame 4: The person is sitting and working at the sewing machine.
Frame 5: A person is sitting and crafting, possibly sewing or working with fabric.
Frame 6: The person is sitting. The person is holding a small object. The person is crafting or creating something, as indicated by the presence of a sewing machine and related crafting supplies.
…

Verbs:
Frame 1: Someone is creating. Someone is using a tool. Someone is using a tool. Someone is using a tool. Someone is using a tool. Someone is applying a substance. Someone is cleaning up. Someone is possibly consuming.
Frame 2: Someone is sitting and working. Someone is using something. Someone is cutting.
Frame 3: Someone is sitting and using something. Someone is working on something. Something is being stored. Someone is possibly cutting something.
Frame 4: Someone is working.
Frame 5: Someone is sitting and working.
Frame 6: Someone is sitting. Someone is holding something. Someone is creating something.
…

Figure F.4. Qualitative examples of objects, activities, and verbs textual descriptions for random videos in MSRVTT dataset.

(a) Examples from the UCF101 test split (that are not included in the UCF101-UTD split)

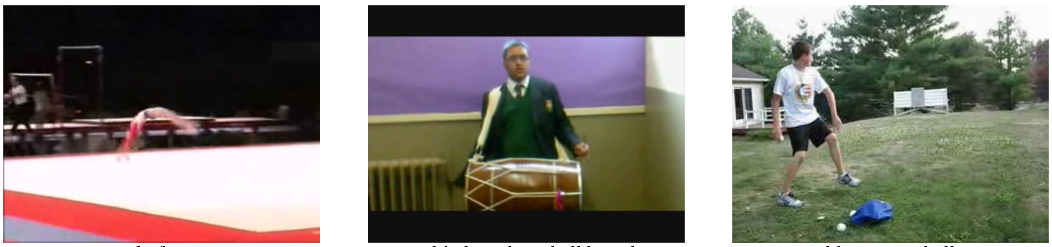


(b) Examples from our UCF101-UTD-split

Figure F.5. Video examples with their class label from full UCF101 test set and our object-debiased UTD-split. We observe that samples from our object-debiased UTD-split require a level of video understanding beyond simple object recognition. For instance, in the case of playing musical instruments, e.g., Playing Daf or Playing Cello, the videos often include other musical instruments in the background, e.g., a piano or drums in the case of Playing Daf, alongside the primary instrument. Similarly, in the Pizza Tossing class, the pizza in the UTD-split example is hardly visible, and a video requires analysis beyond this single frame for correct class prediction.

sleeveless top | snare drum | upright piano

(a) Attribute error



platform | whiteboard or chalkboard | blue sports ball

(b) Misclassification

rows of seats | metallic base | fence

(c) Hallucination

white rails | white-capped waves | red strap

(d) Human annotation mistake – the object is visible

body of water | air | large

(e) Other

Figure F.6. Examples of manual classification of objects predicted by the VLM for the image, but not selected as visible in the image during the user study. We consider five categories: attribute error, misclassification, hallucination, human annotation mistake (the object is visible), and other.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017. 8, 10

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 6

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 2

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 8, 10

[5] João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arxiv*, 1808.01340, 2018. 6, 8, 10

[6] João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arxiv*, 1907.06987, 2019. 6, 8, 10

[7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts. In *CVPR*, 2021. 6

[8] Pradipto Das, Chenliang Xu, Richard F. Doell, and Jason J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*, 2013. 8, 10

[9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021. 1

[10] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The ”something something” video database for learning and evaluating visual common sense. In *ICCV*, 2017. 6, 8, 10

[11] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 6

[12] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 1, 6

[13] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arxiv*, 1705.06950, 2017. 6, 8, 10

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*. 6

[15] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022. 6

[16] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *ICCV*, pages 19948–19960, 2023. 5, 6

[17] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*, 2024. 5, 6

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 6

[20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 2, 4

[21] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 2, 4, 6

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[23] Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. Sfrembedding-mistral: enhance text retrieval with transfer learning. *Salesforce AI Research Blog*, 3, 2024. 1, 4

[24] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 6

[25] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *TPAMI*, 42(2):502–508, 2019. 8, 10

[26] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken moments: Learning joint audio-visual representations from video descriptions. In *CVPR*, 2021. 8, 10

[27] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *NeurIPS*, 24, 2011. 6

[28] 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. 1, 2, 4

[29] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, 2015. 8, 10

[30] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018. 6

[31] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arxiv*, 1705.06950, 2017. 8, 10

[32] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*, 35:10078–10093, 2022. 5, 6

[33] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *CVPR*, pages 6598–6608, 2023. 6

[34] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *CVPR*, pages 14549–14560, 2023. 5, 6

[35] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2023. 1, 4

[36] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*, 2023. 6

[37] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 8, 10

[38] Liangzhe Yuan, Nitesh Bharadwaj Gundavarapu, Long Zhao, Hao Zhou, Yin Cui, Lu Jiang, Xuan Yang, Menglin Jia, Tobias Weyand, Luke Friedman, et al. Videoglue: Video general understanding evaluation of foundation models. *arXiv preprint arXiv:2307.03166*, 2023. 5

[39] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *NeurIPS*, 2021. 6

[40] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *ECCV*, pages 310–325. Springer, 2025. 1, 4