Contents

A. Contrastive pretraining for language-aware models

We explain the surge of computational overhead for language-aware models under contrastive pretraining. Since goal-free perception independently encodes vision and language, for B video-language pairs, we only need to encode B video clips and then compute a similarity matrix with a shape of $B \times B$. However, our experiments confirm VideoDistill will fast degenerate if we just contrast between matched pairs (calculate a single representation for each video based on its matched annotation and compute a $B \times B$ similarity matrix as we do in Equation 7). The reason for this phenomenon is the video encoder simultaneously takes matched video-language pairs as input. It can simply meet the requirements of the contrastive objectives if its output is always identical with language inputs, whatever video is received. To avoid degeneration, the comparison can not be limited to matched pairs. We should encode videos with all possible annotations in the mini-batch (compute B^2 video representations and $B \times B^2$ similarity matrix). Also, we should constrain each video representation based on an unmatched annotation to be unfamiliar with the videos' matched annotations.

Nevertheless, the full contrastive learning for language-aware models leads to a quadratic growth in computational overhead. This demand is beyond the reach of our current resources. We will further study this full contrastive learning in future work.

B. Training Details

B.1. Pretraining Details

Pretraining Datasets.b Our pretraining set consists of three parts: (1) 3M video-caption pairs randomly sampled from generic dataset WebVid10M [3]. (2) 4.2M video-caption pairs randomly sampled from YouTube video dataset HD-VILA [31]. We ensure the lengths of video clips sampled from WebVid10M and HD-VILA range from 10s to 30s. (3) 3.8M video-caption pairs from the 1st-person view dataset EgoCLIP [19]. Generally speaking, the 1st-person videos have more significant changes in perspective and orientation as the user moves around than the 3rd-person videos. Thus, they are helpful in releasing the potential of solving multiple events and multi-scale reasoning for VideoDistill.

Table 1. Comparison with SOTA methods on MSRVTT-QA.

Method	Pretraining data	Pairs	Acc
ST-VQA[9]	-	-	30.9
Co-Memory [7]	-	-	32.0
AMU [28]	-	-	32.5
HME [6]	-	-	33.0
SSML [1]	HowTo100M [22]	136M	35.1
HCRN [14]	-	-	35.6
ClipBert [16]	COCO [4],VisGenome [13]	2.1M	37.4
CoMVT [26]	HowTo100M [22]	136M	39.5
HD-VILA [31]	HD-VILA-100M [31]	100M	40.0
PMT [23]	-	-	40.3
VQA_T [32]	-	-	39.6
VQA_T [32]	HowToVQA69M [32]	69M	41.5
ALPRO [17]	HowTo,WebVid	5.5M	42.1
VideoDistill†	-	-	42.7
VideoDistill	WebVid,HD-VILA,EgoCLIP	11M	**44.2**

Table 2. Comparison with SOTA methods on MSVD-QA.

Method	Pretraining set	Pairs	Acc
HME [6]	-	-	33.7
SSML [1]	HowTo100M	136M	35.1
HCRN [14]	-	-	36.1
PMT [23]	-	-	41.8
CoMVT [26]	HowTo100M	136M	42.6
SiaSamRea [33]	COCO,VisGenome	2.1M	45.5
ALPRO [17]	HowTo,WebVid	5.5M	45.9
VQA_T [32]	HowToVQA69M	69M	46.3
VideoDistill†	-	-	46.2
VideoDistill	WebVid,HD-VILA,EgoCLIP	11M	**49.2**

Implementation Details We resize all video clips (as well as downstream videos) to 256p while preserving the aspect ratio, then extract frames with 7.5 fps. We randomly sample 100 frames as input during pretraining and evenly sample 100 frames for downstream tasks. Finally, we augment input frames by random crop a 224×224 region to increase input diversity.

In the video branch, we adopt CLIP-ViTB/16 [25] as the frame encoder. FS-Blocks and VB-Blocks have $L = 3$ layers, a hidden size of $D = 1024$. The number of attention heads equals 8 for all LA-Gates, self-attention layers, and spatial-temporal layers. We borrow spatial-temporal layers from FrozenInTime [3]. We add a learnable temporal embedding for the input of the first FS-Block, a learnable temporal embedding, and a spatial embedding for the input of the first vision refinement block. We sparsely sample $K = 16$ frames from 100 densely sampled frames as the input of vision refinement blocks for most experiments unless otherwise specified. In the text branch, we utilize the text encoder from CLIP with a maximum sequence length of 77.

For all experiments, we use AdamW optimizer with a

Table 3. Results on EgoMCQ multiple-choice test.

Methods	Pretraining set	Pairs	Intra-video ACC(%)	Inter-video ACC(%)
TimeSFormer+Distillbert	EgoCLIP	3.8M	85.5	47.0
FrozenInTime [3]	EgoCLIP	3.8M	89.4	51.5
EgoNCE w/Pos [19]	EgoCLIP	3.8M	89.7	53.6
EgoNCE w/Pos&Neg [19]	EgoCLIP	3.8M	90.6	57.2
EgoVLP-v2 [24]	EgoCLIP	3.8M	91.0	60.9
VideoDistill†	-	-	92.0	59.0
VideoDistill	WebVid,HD-VILA,EgoCLIP	11M	**92.7**	**61.3**

Table 4. Results on MSRVTT-multiple-choice test.

Method	Pretraining set	Pairs	Acc
CT-SAN [34]	-	-	66.4
MLB [12]	-	-	76.1
JSFusion [35]	-	-	83.4
ActBERT [37]	HowTo100M	-	85.7
ClipBert [16]	COCO,VisGenome	2.1M	88.2
VideoCLIP[29]	HowTo100M	136M	92.1
HD-VILA [31]	HD-VILA-100M	100M	97.1
VideoDistill	WebVid,HD-VILA,EgoCLIP	11M	**97.8**

learning rate of 3×10^{-5} and a weight decay of 1×10^{-3}. Also, we employ a linear decay learning rate schedule with a warm-up strategy. We pretrain VideoDistill on 8 A100 GPUs with a batch size of 256 for 2 epochs (53 hours) to get our model applied to downstream tasks. Note that downstream performances may be further improved if we train the model for more epochs or customize better hyperparameters of the model architecture.

Table 5. Comparison of text-to-video retrieval on MSR-VTT, 1k-A split. †denotes our model finetuned in a contrastive manner.

Method	PT-set	PT-pairs	R@1	R@5	R@10
CE[20]	-	-	20.9	48.8	62.4
UniVL[21]	HowTo100M	136M	21.2	49.6	63.1
ClipBERT[16]	COCO,VisGenome	5.6M	22.0	46.8	59.9
FrozenInTime[3]	CC3M,WV2M,COCO	6.1M	32.5	61.5	71.2
VideoCLIP[29]	HowTo100M	136M	30.9	55.4	66.8
HD-VILA [31]	HD-VILA-100M	100M	**35.6**	65.3	**78.0**
VideoDistill†	WebVid,HD-VILA,EgoCLIP	11M	32.8	63.5	<u>74.0</u>
VideoDistill	WebVid,HD-VILA,EgoCLIP	11M	<u>33.4</u>	**70.1**	72.9

Table 6. Comparison of text-to-video retrieval on DiDeMo. †denotes generating the results of retrieval by direct similarity comparison like previous works (otherwise, by VTM head) during fine-tuning.

Method	PT-set	PT-pairs	R@1	R@5	R@10
HERO[18]	TV[15],HowTo	7.6M	2.1	-	11.4
S2VT[27]	COCO	-	11.9	33.6	-
FSE [36]	Sports-1M[11]	1M	13.9	36.0	-
CE[20]	-	-	16.1	46.1	-
ClipBERT[16]	COCO,VisGenome	5.6M	20.4	48.0	60.8
HD-VILA [31]	HD-VILA-100M	100M	**28.8**	<u>57.4</u>	**69.1**
VideoDistill†	WebVid,HD-VILA,EgoCLIP	11M	<u>28.0</u>	57.1	<u>66.4</u>
VideoDistill	WebVid,HD-VILA,EgoCLIP	11M	27.2	**61.6**	63.1

B.2. Finetuning Details

Finetuning Datasets.

EgoMCQ [19] is a 1st-person Multiple-Choice Questions answering task. Each text query has five video candidates. It provides two criteria named Inter-video and intra-video accuracy. The former ensures the five video candidates come from different videos, and the latter collects candidates from the same video. The evaluation metric is accuracy.

MSRVTT-QA [28] and **MSRVTT-multiple-choice test** [35] are two video question answering tasks basd on MSRVTT [30]. The former is open-ended, and the latter is multiple-choice. The evaluation metric is accuracy.

MSVD-QA [28] is an open-ended question answering task with 1.9k short generic video clips. The evaluation metric is accuracy.

EgoTaskQA [10] is a long-form open-ended dataset with an average video length of 25s. It provides 15 categories of questions to evaluate models in detail. It also provides a version of the dataset (*indirect* split) to reduce the usage of language shortcuts. The evaluation metric is accuracy.

AGQA [8] a long-form open-ended dataset contains 8 types of compositional spatiotemporal reasoning. The average video length is 30s. We use its v2 version, which has more balanced distributions, as the dataset creator recommended. The evaluation metric is accuracy.

MSRVTT [30] is 3rd-person video-text retrieval task. It contains 10K YouTube videos. We follow previous works [31, 35], finetuning SpaceCLIP on 9K videos and reporting results on the 1K-A test set. The evaluation metric is **R**@1,

Table 7. Language-only QA results on the EgoTaskQA *normal* split. (Gussian inputs)

Category	VisualBERT [5]		HCRN (w/o vision)		**VideoDistill (w/o vision)**	
	Acc.	Change	Acc.	Change	Acc.	Change
world	36.28	-8.7%	35.22	-20.4%	32.06	-32.2%
intent	35.02	-21.3%	34.93	-29.8%	26.56	-49.4%
multi-agent	20.58	-21.7%	19.17	-38.9%	18.58	-49.7%
descriptive	34.55	-17.7%	33.58	-22.8%	29.45	-36.7%
predictive	24.75	-18.5%	24.3	-33.5%	19.93	-50.7%
counterfactual	41.3	-1.6%	40.4	-15.8%	39.51	-20.4%
explanatory	31.78	-15.1%	30.57	-24.7%	26.84	-36.9%
action	15.72	+4.6%	15.64	-1.7%	15.93	-2.6%
object	7.43	-68%	6.33	-86.0%	2.68	-95.1%
state	45.03	-23.9%	42.51	-37.7%	33.33	-53.9%
change	69.87	+2.3%	68.77	+2.1%	63.67	-10.9%
all	33.92	-10.6%	32.51	-23.0%	29.45	-33.9%

Table 8. Performances on the EgoTaskQA *indirect* split.

	Category	BERT	HCRN (w/o vision)	VisualBERT	PSAC	HME	HGA	HCRN	ClipBERT	**VideoDistill†**
Scope	world	34.96	33.61	40.00	44.74	35.91	31.29	44.04	26.51	**47.82**
	intent	23.56	23.98	36.02	48.38	31.73	20.42	47.02	14.66	**49.61**
	multi-agent	19.70	19.25	26.02	**35.37**	25.07	17.74	30.11	20.09	<u>35.04</u>
Type	descriptive	33.09	30.73	38.9	43.36	34.48	29.01	42.02	24.35	**45.13**
	predictive	15.58	13.68	31.37	29.11	27.79	15.16	46.32	10.32	**52.83**
	counterfactual	34.59	34.75	37.63	39.94	35.07	33.01	43.64	26.29	**43.97**
	explanatory	27.38	28.11	32.75	42.53	29.16	24.00	39.69	22.46	**43.75**
Semantic	action	26.91	28.18	27.49	30.06	25.12	26.15	29.61	25.25	**30.34**
	object	2.808	4.13	22.63	30.97	19.08	7.02	32.20	10.49	**45.97**
	state	21.96	21.24	32.02	43.29	31.60	17.67	41.81	15.29	**49.77**
	change	55.28	50.71	55.59	**57.20**	47.65	47.22	56.27	35.26	<u>53.98</u>
	all	31.78	30.76	37.01	42.25	33.06	28.36	41.56	24.08	**44.77**
Performance Change		6.4%	5.4%	2.4%	4.9%	17.7%	22.9%	1.5%	39.6%	**0.25%**

R@5, **R**@10.

DiDeMo [2] consists of 10K Flickr videos and 40K manually annotated sentences. We use a standard split to fine-tune VideoDistill on the training set and report the result on the test set. The evaluation metric is **R**@1, **R**@5, **R**@10.

Implementation Details. For open-ended datasets MSRVTT-QA and MSVD-QA, EgoTaskQA, and AGQA, we take questions as the language input, then encode the answers in a one-hot fashion and train a two-layer MLP classification head over all answer candidates with a cross-entropy loss on the top of visual representation v^*_{cls}. For the multiple-choice dataset EgoMCQ, we respectively combine the five candidate videos with the question to form five input pairs, then choose the video corresponding with the maximum logit over the VTM head as the answer. For the multiple-choice dataset MSRVTT-multiple-choice test, we concatenate five answers with the question into five sentences, then choose the answer with the maximum logit over the VTM head. For text-to-video retrieval MSRVTT and DiDeMo, we provide two ways to realize retrieval. The first method is finetuning the module in a contrastive manner and choosing the answer with the highest similarity of v^*_{cls} and t_{cls}. The second method is choosing the answer with the highest VTM logits. We set the batch size to 128 and finetune the pretrained VideoDistill on 4 A100 GPUs.

C. Generic VideoQA

We evaluate VideoDistill on the four commonly used VideoQA datasets: **MSRVTT-QA** [28], **MSVD-QA** [28], **EgoMCQ** [19] and **MSRVTT-multiple-choice test** [35].
Results. In Table 1,2,3,4, the result of VideoDistill shows that our model outperforms existing methods on four tasks. On open-ended datasets MSRVTT-QA and MSVD-QA, we achieve 2.1% and 2.9% improvement over SOTA methods. Especially our from-scratch model outperforms previous large-scale pretrained models with 0.6% gains. For multiple-choice datasets EgoMCQ and MSRVTT-multiple-choice test, the task setting is more like the retrieval and is more suitable for contrastive frameworks like HD-VILA[31] and VideoCLIP[29]. Our model is still better than the SOTA methods. We find that VideoDistill achieves an improvement of 2.1% on EgoMCQ Intra-video test, which is challenging since it ensures the five candidate answers are continuous clips with similar visual appearances. It shows that VideoDistill can better extract question-related visual semantics.

D. Video-Text Retrieval

Although VideoDistill is specially designed for VideoQA, we still evaluate it on text-to-video retrieval datasets MSRVTT [30] and DiDeMo [2] to show its generalization power in Table 5 and Table 6.

E. More quantitative results and ablations

The impact of LA-Gate. To further demonstrate that LA-Gate can reduce the use of language prior, we eport the performance degradations of replacing visual inputs with Gaussian noise in Table 7. Similar to section 4.4 Table 3, we find that VideoDistill relies more on visual reasoning during the answer generation.

We also test VideoDistill on EgoTaskQA indirect split, which is motivated by the fact [10] that during task execution, actions, objects, and their changes are often strongly correlated. It leaves the chance for the model to perform well by simply over-fitting these strong correlations (language bias) without thorough task understanding. The indirect references can avoid these correlations. Table 8 shows that our VideoDistill has the least absolute performance change. It indicates that VideoDistill barely utilizes language bias in questions.

The choice of the number of densely sampled frames. We conduct the experiments in Table 9 with $L = 3$ and 16 encoded frames. We find that longer video clips (Ego-taskQA) require a larger N to ensure we are not omitting the necessary information. Nevertheless, too large N will damage the performance. One possible reason is a larger N needs more stacked frame sampling blocks. However, larger L consumes more computing resources.

Reasonable number of stacked layers L. In Table 10, we set $N = 100$ and simultaneously change L for differentiable sparse sampling and vision refinement. We find too many layers still damage the performance since bigger L dramatically improve the models' ability of fitting. Models will easily trapped in local minimums.

The effectiveness of pretraining losses. The designing concepts of pretraining losses are: MLM improves context reasoning by predicting the masked token. VTM and CL align visual and textual embeddings. Most of the time, applying one of VTM and CL is enough. This paper utilizes an incomplete CL to stabilize the training. Ablations on pretraining loss are shown in Table 11.

Table 9. Sensitivity to densely sampled frames.

N	EgoTaskQA	MSRVTT-QA
50	40.86	42.13
100	**45.02**	**44.20**
150	44.80	42.15
200	42.12	41.10

Table 10. Sensitivity to the number of stacked blocks.

L	EgoTaskQA	MSRVTT-QA
1	35.50	24.85
3	**45.02**	**44.20**
5	43.60	44.1
8	42.18	43.59

Table 11. Analysis of the effectiveness of pretraining tasks.

F. Qualitative Results

We visualize the result of our differentiable sparse sampling module. Specifically, we report two instances from a four-frame variant (the number of selected frames $K = 4$) in Figure 1 and a full instance from the sixteen-frame version used on downstream tasks in Figure 2. Note that models with $K > 4$ allow duplicate selection, which means important frames can appear more than once in the K selected frames.

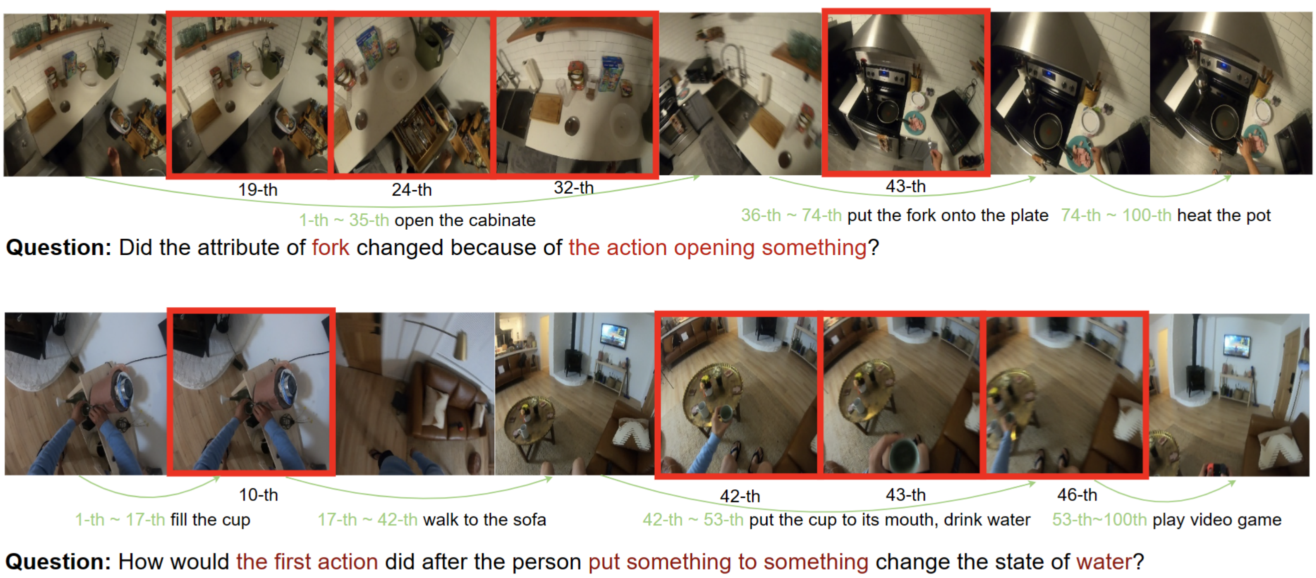

Figure 1. Two instances from the four-frame variant

Question: What is the person doing after he/she close something?
Answer: open bottle-water

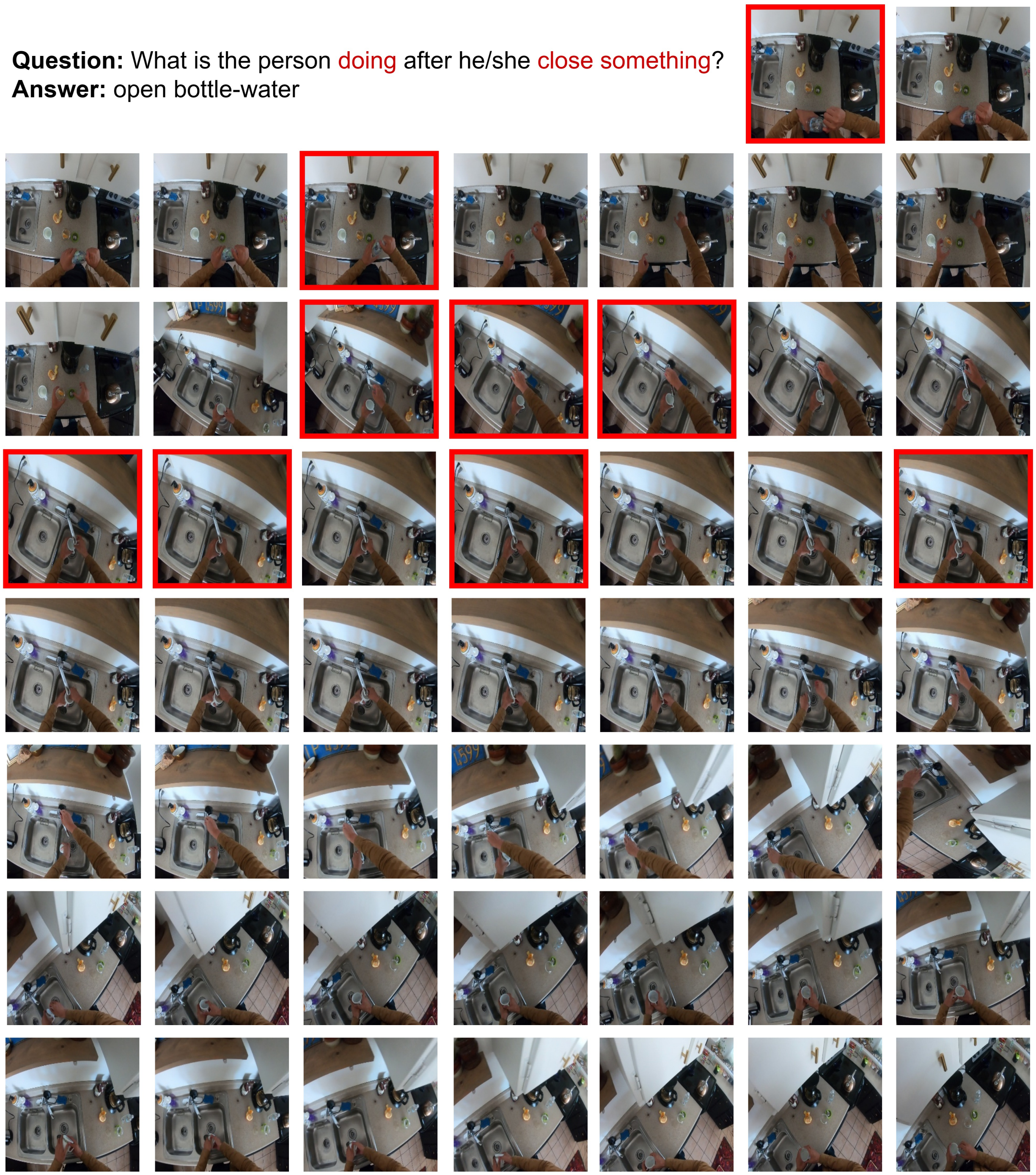

Figure 2. A full instance from the 16-frame variant

References

[1] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6644–6652, 2021. 1

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 3, 4

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 1, 2

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[6] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 1

[7] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6576–6585, 2018. 1

[8] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 2

[9] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 1

[10] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *arXiv preprint arXiv:2210.03929*, 2022. 2, 4

[11] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2

[12] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016. 2

[13] 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. *International journal of computer vision*, 123:32–73, 2017. 1

[14] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981, 2020. 1

[15] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 2

[16] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 1, 2

[17] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022. 1

[18] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 2

[19] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022. 1, 2, 4

[20] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 2

[21] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 2

[22] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 1

[23] Min Peng, Chongyang Wang, Yu Shi, and Xiang-Dong Zhou. Efficient end-to-end video question answering with pyramidal multimodal transformer. *arXiv preprint arXiv:2302.02136*, 2023. 1

[24] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5285–5297, 2023. 2

[25] 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

[26] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16877–16887, 2021. 1

[27] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014. 2

[28] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 1, 2, 4

[29] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 2, 4

[30] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2, 4

[31] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022. 1, 2, 4

[32] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. 1

[33] Weijiang Yu, Haoteng Zheng, Mengfei Li, Lei Ji, Lijun Wu, Nong Xiao, and Nan Duan. Learning from inside: Self-driven siamese sampling and reasoning for video question answering. *Advances in Neural Information Processing Systems*, 34: 26462–26474, 2021. 1

[34] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3165–3173, 2017. 2

[35] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 2, 4

[36] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the european conference on computer vision (ECCV)*, pages 374–390, 2018. 2

[37] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 2