Appendix

In Appendix A we provide additional details on how the different tasks of Ego4D [25] are modelled as a temporal graph, expanding Sec. 4.1 of the main paper. Appendix B provides additional implementation details of the *Task Translation* model and how the EgoT2 architecture was adapted to our scenario. Appendix C presents additional experiments to evaluate the role of negative transfer in MTL. A more in-depth comparison of the methods on the test-set of the Ego4D challenges is shown in Appendix D. Finally, we show more qualitative results in Appendix E.

A. Additional implementation details

A.1. Temporal modelling

Action Recognition (AR) Action annotations are derived from the LTA benchmark. Each action is an annotated segment lasting approx. 8.0 seconds, and actions may temporally overlap. To provide each action with additional temporal context from the surrounding, we process actions in fixed length sequences of $w = 9$ actions, each mapped to a node in the temporal graph. The target action is the central node of the sequence and the classification loss is computed only on this node. The window size was selected to be larger than the receptive field of the temporal GNN after 3 layers of graph convolution, which we observed to be the optimal depth of the network. Furthermore, using fixed size sequences allows to train the model on videos containing a variable and possibly large number of actions.

Long Term Anticipation (LTA) LTA is formulated as an action anticipation task in which the model is shown $N = 2$ input clips and has to predict the actions occurring in the following $Z = 20$ timestamps. As in AR, clips lasts approx. 8.0 seconds and may overlap. Therefore, the effective temporal window seen by the model may vary between 8.0 and 16.0 seconds, depending on how much the input clips overlap. Input clips and future timestamps are mapped to nodes in the graph, with the latter initialised with the mean of the features of the input clips. Edges connect each node to its subsequent and preceding nodes. Additionally, nodes that represent future actions to be predicted are also connected to the input clips. This connectivity pattern allows local temporal reasoning, *e.g.* to rearrange the order of actions in the anticipation window, while using the global context provided by the input clips to guide the prediction. Similarly to AR, there is a *one-to-one* correspondence between actions and nodes of the graph.

Object State Change Classification (OSCC) and Point of No Return (PNR) Unlike action-based annotations, OSCC and PNR do not necessarily match the boundaries of an action segment of the video. Each segment lasts approx. 8.0 seconds and is uniformly divided in 4 (OSCC) or 16 (PNR) smaller sub-segments that are mapped to the nodes of the graph.

A.1.1 Temporal model sharing across tasks

A key premise of EgoPack is that different tasks are modelled using the same shared temporal backbone architecture, even though the temporal granularity of the different tasks may vary. To achieve this, we do not constraint nodes to represent the same fixed size temporal window across all tasks. Through the utilisation of a multi-task learning process, we force the network to jointly learn tasks with different temporal resolutions, enabling reasoning at different temporal scales. This formulation is particularly effective to prepare the model to new tasks, as the model has already learnt to combine tasks with potentially different temporal resolutions during the MTL training. As an example, consider the case in which the MTL model is trained on AR, LTA, OSCC and PNR. In this case, the nodes of the temporal graph represent actions when the task is AR or LTA, or shorter temporal sub-segments for OSCC and PNR. To train EgoPack, we update the weights of $\mathcal{M}_t$ for all novel tasks, except LTA for which we observe better performance by not updating the temporal model.

B. Task Translation Implementation Details

The objective of the *Task Translation* experiments is to compare the task translation mechanism proposed by EgoT2 [68], which learns a mapping between features extracted from different task-specific models, to EgoPack which leverages past gained knowledge under the form of task-specific prototypes. For fair comparison, we re-implement this mechanism and evaluate it on top of the same temporal backbone and the same pre-extracted features of EgoPack. We start from the EgoT2-g model and employ the same architecture for the *Task Translation*, which consists of a 1-layer encoder-decoder stack, each with 8 heads, dropout 0.1 and features size 1024. The input of the *Task Translation* is provided by an ensemble of *Temporal Graph* models, one for each task. The whole architecture is trained for one task at the time, as EgoPack, and only the encoder-decoder architecture is updated, while the temporal models that compose the ensemble are kept frozen. We train *Task Translation* for 30 epochs, using the Adam optimiser with learning rate 1×10^{-4} (with the exception of OSCC which uses learning rate 1×10^{-3}), batch size 16, linear warmup for the first 5 epochs and weight decay 1×10^{-5}.

C. Additional Multi-Task Experiments

MTL suffers from negative transfers between different tasks, and fine-tuning an MTL on a new task may not be

	AR		OSCC	LTA		PNR
	Verbs Top-1 (%)	Nouns Top-1 (%)	Acc. (%)	Verbs ED (↓)	Nouns ED (↓)	Loc. Err. (s) (↓)
Temporal Graph	**24.25**	**30.43**	**71.26**	0.754	0.752	**0.61**
Multi-Task Learning	22.16	29.34	70.93	0.740	**0.746**	0.62
Multi-Task Learning (+ PCGrad [71])	22.01	29.46	70.86	**0.737**	**0.746**	0.63

Table 4. Results of PCGrad [71] compared to vanilla Multi-Task Learning.

PNR	Pre-trained on Ego4D [25]	Trained on pre-extracted features	Loc. Error (s) (↓)		
CNN LSTM [25]	✗	✗	0.76		
EgoVLP [42]	✓	✗	0.67		
EgoT2 [68]	✗	✗	**0.66**		
EgoPack	✗	✓	**0.66**		
OSCC	**Pre-trained on Ego4D [25]**	**Trained on pre-extracted features**	**Accuracy (%)**		
I3D RN-50 [25]	✗	✗	67.6		
EgoVLP [42]	✓	✗	74.0		
EgoT2 (EgoVLP) [68]	✓	✗	**75.0**		
EgoT2 (I3D) [68]	✗	✗	71.0		
EgoPack (SlowFast)	✗	✓	72.1		
LTA	**Pre-trained on Ego4D [25]**	**Trained on pre-extracted features**	**Verb (↓)**	**Noun (↓)**	**Action (↓)**
SlowFast [25]	✗	✗	0.739	0.780	0.943
EgoT2 [68]	✗	✗	0.722	0.764	0.935
HierVL [1]	✓	✗	0.724	**0.735**	0.928
I-CVAE [43]	✗	✓	0.741	0.740	0.930
EgoPack	✗	✓	**0.721**	**0.735**	**0.925**

Table 5. Comparison of EgoPack on the test set of the Ego4D benchmarks, highlighting differences in terms of additional Ego4D pretraining and use of pre-extracted features.

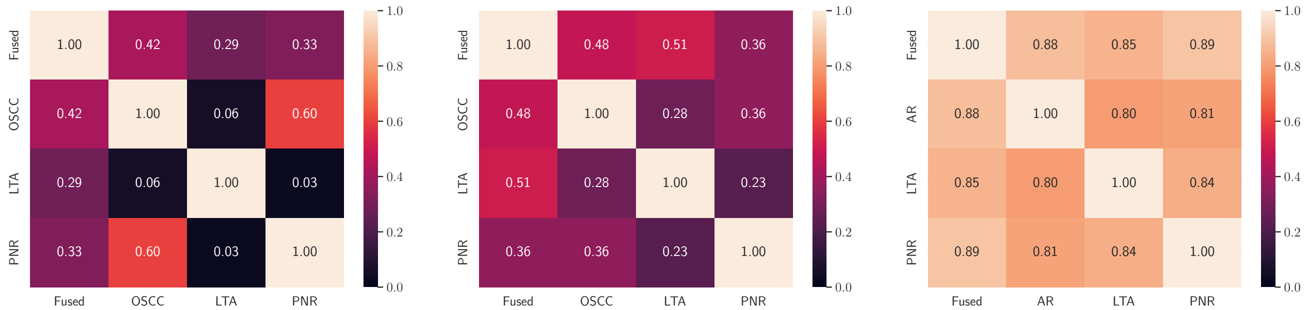
the most effective approach to retain knowledge learned in the MTL training process. We observe evidence of this phenomenon in Table 6, where we compare the MTL on all tasks with two finetuning approaches to extend a model trained on three tasks to a fourth novel task. MTL+FT finetunes the model for the novel task, as already shown in Table 2, while MTL+TT replaces the EgoPack's second stage with a decoder analogous to TT, which learns the new task as a "recombination" of the previous tasks.

Brute Force Multi-Task Learning Table 7 presents a comprehensive analysis of MTL on all task combinations, to assess the effect of negative transfer when a smaller subset of tasks is used. Even with fewer tasks, MTL still suffers from negative transfer across tasks and does not represent an upper bound for EgoPack, which is showing a clear advantage.

Minimising negative transfer Various approaches have been proposed to address the issue of negative transfer in multi-task learning [9, 27, 38, 59, 63, 71]. Although the multi-task setting significantly differs from the settings proposed for EgoPack, we provide a comparison with one of these methods, PCGrad [71], which projects tasks' gradients on the normal plane of all the other gradients to remove interference among tasks. Apart from minimal fluctuations, PCGrad does not appear to significantly improve over MTL, showing that these methods may still be insufficient to effectively reduce the negative transfer, as shown in Table 4.

D. Comparison of methods on the test-set

We summarise the main differences between EgoPack and the other methods on the test-set in Table 5, extending Table 3 of the main paper and highlighting differences in terms of additional Ego4D pretraining and use of pre-extracted features. EgoPack relies on pre-extracted features from Omnivore [24], which was trained on Kinetics-400 [4] for action recognition. As a result, these features are highly semantic and may struggle to encode finer temporal details required by certain tasks, *e.g.* to detect changes in the objects being manipulated in OSCC or PNR. Most other methods, with the exception of I-CVAE [43], train also their features



(a) AR Verb (b) AR Noun (c) OSCC

Figure 7. Agreement ratio between predictions from different tasks when the novel task is Action Recognition (Fig. 7a and Fig. 7b) and Object State Change Classification (Fig. 7c). *Fused* represents the sum of the logits from the auxiliary tasks.

	AR		OSCC	LTA		PNR
	Verbs Top-1 (%)	**Nouns Top-1 (%)**	**Acc. (%)**	**Verbs ED (↓)**	**Nouns ED (↓)**	**Loc. Err. (s) (↓)**
MTL (All tasks)	22.05	29.44	71.10	0.740	**0.746**	0.62
MTL (3 tasks) + FT	24.36	**31.31**	71.60	0.744	0.754	0.62
MTL (3 tasks) + TT	22.30	29.50	70.96	0.738	0.757	0.62
EgoPack	**25.10**	31.10	**71.83**	**0.728**	0.752	**0.61**

Table 6. Comparison of vanilla MTL and two finetuning strategies to extend MTL models to novel tasks.

extraction backbones on Ego4D benchmarks' data, which allows to learn task-specific models more suited for the task at hand. On the contrary, we do not update the features extraction backbone when training EgoPack.

When evaluating EgoPack on the test-set, we also observe a significant performance gap compared to other methods that rely on some amount of additional data from Ego4D, while the benchmarks data are more limited in size. HierVL [1] is pretrained on the full Ego4D using a contrastive video-language objective with short-term and long-term textual narrations. EgoVLP [42] is pretrained on a large subset of Ego4D using a video-language contrastive objective with action-aware positive samples and scene-aware negative samples. The only method directly comparable to EgoPack in terms of pre-training data and parameters updated is I-CVAE [43], which uses the SlowFast [16] features released by [25] for the LTA benchmark. The extension of EgoPack to additional backbones, possibly with end-to-end finetuning, is outside of the scope of this paper and is left as a future work. For OSCC, we report the results of EgoPack using SlowFast features instead of Omnivore as they showed better performances compared to the latter.

E. Additional qualitative results

EgoPack fuses the predictions coming from different task perspectives by summing the task-specific logits. We show in Fig. 7 the agreement ratio between the predictions produced by the different tasks $\mathbf{y}_i^k$ and the final output computed as the sum of the individual contributions $\mathbf{y}_i = \sum_k \mathbf{y}_i^k$. In Action Recognition, we observe low agreement both between task pairs and with respect to the fused predictions, suggesting that they contribute complementary information to the novel task. On the other hand, in OSCC, tasks predictions tend to be more consistent across tasks.

References

[1] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *CVPR*, 2023. 2, 8, 10, 11

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[3] Alejandro Betancourt, Pietro Morerio, Carlo S Regazzoni, and Matthias Rauterberg. The evolution of first person vision methods: A survey. *IEEE TCSVT*, 2015. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 5, 10

[5] Rich Caruana. Multitask learning. *Machine learning*, 28: 41–75, 1997. 3

[6] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *ICCV*, 2019. 2

[7] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. In *NeurIPS*, 2022. 3

AR	OSCC	LTA	PNR	AR		OSCC	LTA		PNR
				Verbs Top-1 (%)	Nouns Top-1 (%)	Acc. (%)	Verbs ED (↓)	Nouns ED (↓)	Loc. Err. (s) (↓)
Single tasks				24.25	30.43	71.26	0.754	0.752	**0.61**
✓	✓	-	-	23.98	30.60	70.81	-	-	-
✓	-	✓	-	22.23	29.48	-	0.744	0.744	-
✓	-	-	✓	24.05	30.72	-	-	-	0.63
-	✓	✓	-	-	-	70.71	0.745	0.751	-
-	✓	-	✓	-	-	71.01	-	-	0.66
-	-	✓	✓	-	-	-	0.751	0.752	0.62
✓	✓	✓	-	22.05	29.44	71.10	0.739	0.745	-
✓	✓	-	✓	23.82	30.83	71.03	-	-	0.63
✓	-	✓	✓	22.24	29.83	-	0.745	**0.743**	0.62
-	✓	✓	✓	-	-	71.06	0.746	0.751	0.63
MTL (All tasks)				22.05	29.44	71.10	0.740	0.746	0.62
EgoPack				**25.10**	**31.10**	**71.83**	**0.728**	0.752	**0.61**

Table 7. Brute force experiments in multi-task learning with all combinations of tasks.

[8] Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, and Yeqing Li. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In *ICCV*, 2023. 3

[9] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018. 3, 10

[10] Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and Wanli Ouyang. Unihcp: A unified model for human-centric perceptions. In *CVPR*, 2023. 3

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE TPAMI*, 2021. 2

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *IJCV*, 2022. 2

[13] Eadom Dessalene, Michael Maynord, Chinmaya Devaraj, Cornelia Fermuller, and Yiannis Aloimonos. Egocentric object manipulation graphs. *arXiv preprint arXiv:2006.03201*, 2020. 3

[14] Eadom Dessalene, Chinmaya Devaraj, Michael Maynord, Cornelia Fermuller, and Yiannis Aloimonos. Forecasting action through contact representations from first person video. *IEEE TPAMI*, 2021. 3

[15] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation. In *The world wide web conference*, 2019. 2

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 3, 11

[17] Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. In *NeurIPS*, 2021. 3

[18] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE TPAMI*, 2020. 2

[19] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 2017. 2

[20] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *CVPR*, 2020. 2

[21] Pallabi Ghosh, Nirat Saini, Larry S Davis, and Abhinav Shrivastava. All about knowledge graphs for actions. *arXiv preprint arXiv:2008.12432*, 2020. 3

[22] Pallabi Ghosh, Yi Yao, Larry Davis, and Ajay Divakaran. Stacked spatio-temporal graph convolutional networks for action segmentation. In *WACV*, 2020. 3

[23] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *ICCV*, 2021. 2

[24] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *CVPR*, 2022. 3, 5, 10

[25] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 2, 5, 7, 8, 9, 10, 11

[26] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *PR*, 2018. 2

[27] Michelle Guo, Albert Haque, De-An Huang, Serena Yeung, and Li Fei-Fei. Dynamic task prioritization for multitask learning. In *ECCV*, 2018. 3, 10

[28] Pengsheng Guo, Chen-Yu Lee, and Daniel Ulbricht. Learning to branch for multi-task learning. In *ICML*, 2020. 3
[29] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NeurIPS*, 2017. 4
[30] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE TPAMI*, 2009. 2
[31] Thomas E. Huang, Yifan Liu, Luc Van Gool, and Fisher Yu. Video task decathlon: Unifying image and video tasks in autonomous driving. In *ICCV*, 2023. 3
[32] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE TIP*, 2020. 3
[33] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *CVPR*, 2020. 2, 3
[34] Youngkyoon Jang, Brian Sullivan, Casimir Ludwig, Iain Gilchrist, Dima Damen, and Walterio Mayol-Cuevas. Epic-tent: An egocentric video dataset for camping tent assembly. In *ICCVW*, 2019. 2
[35] Zhuoliang Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *ICML*, 2011. 3
[36] Georgios Kapidis, Ronald Poppe, Elsbeth van Dam, Lucas Noldus, and Remco Veltkamp. Multitask learning to improve egocentric action recognition. In *ICCVW*, 2019. 3
[37] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 2016. 2
[38] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 3, 10
[39] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 2020. 2
[40] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, 2017. 2, 3, 6
[41] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 2021. 2
[42] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. In *NeurIPS*, 2022. 8, 10, 11
[43] Esteve Valls Mascaró, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action anticipation. In *WACV*, 2023. 8, 10, 11
[44] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, 2020. 2
[45] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *CVPR*, 2020. 3
[46] Adrián Núñez-Marcos, Gorka Azkune, and Ignacio Arganda-Carreras. Egocentric vision-based action recognition: A survey. *Neurocomputing*, 2022. 2
[47] Francesca Pistilli and Giuseppe Averta. Graph learning in robotics: a survey. *IEEE Access*, 2023. 2
[48] Mirco Planamente, Chiara Plizzari, Simone Alberto Peirone, Barbara Caputo, and Andrea Bottino. Relative norm alignment for tackling domain shift in deep multi-modal classification. *IJCV*, 2024. 2
[49] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *arXiv preprint arXiv:2308.07123*, 2023. 2
[50] Chiara Plizzari, Toby Perrett, Barbara Caputo, and Dima Damen. What can a cook in italy teach a mechanic in india? action recognition generalisation over scenarios and locations. In *ICCV*, 2023. 2, 5
[51] 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 *ICCV*, 2023. 2
[52] Santhosh K. Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Spotem: Efficient video search for episodic memory. In *ICLR*, 2023. 2
[53] Maheen Rashid, Hedvig Kjellstrom, and Yong Jae Lee. Action graphs: Weakly-supervised action localization with graph convolution networks. In *WACV*, 2020. 3
[54] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017. 3, 7
[55] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *ICML*, 2020. 2
[56] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *CVPR*, 2022. 2
[57] Haosen Shi, Shen Ren, Tianwei Zhang, and Sinno Jialin Pan. Deep multitask learning with progressive parameter sharing. In *ICCV*, 2023. 3
[58] Martin Simonovsky and Nikos Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *CVPR*, 2017. 2
[59] Ayan Sinha, Zhao Chen, Vijay Badrinarayanan, and Andrew Rabinovich. Gradient adversarial training of neural networks. *arXiv preprint arXiv:1806.08028*, 2018. 3, 10
[60] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *ICML*, 2020. 3
[61] Ximeng Sun, Rameswar Panda, Rogerio Feris, and Kate Saenko. Adashare: Learning what to share for efficient deep multi-task learning. In *NeurIPS*, 2020. 3
[62] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recog-

nition from various data modalities: A review. *IEEE TPAMI*, 2023. 2

[63] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *ECCV*, 2020. 3, 10

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[65] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive prototype learning for egocentric action recognition. In *ICCV*, 2021. 3

[66] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics*, 2019. 2

[67] Sen Wu, Hongyang R. Zhang, and Christopher Ré. Understanding and improving information transfer in multi-task learning. In *ICLR*, 2020. 6

[68] Zihui Xue, Yale Song, Kristen Grauman, and Lorenzo Torresani. Egocentric video task translation. In *CVPR*, 2023. 2, 3, 5, 6, 7, 8, 9, 10

[69] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, 2022. 1

[70] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition. In *CVPR*, 2022. 2

[71] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *NeurIPS*, 2020. 3, 10

[72] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *ICCV*, 2019. 3

[73] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *ECCV*, 2022. 1, 2

[74] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34 (12):5586–5609, 2021. 3

[75] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *CVPR*, 2023. 2

[76] Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and Jürgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation. In *WACV*, 2023. 1, 2

[77] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018. 5