Multimodal Transformer Models for Human Action Classification

Zoltán Varga[1], Esteve Valls Mascaro[1], Daniel Jan Sliwowski[1], and Dongheui Lee[1,2]

[1] Technische Universität Wien, Autonomous Systems Lab, Gußhausstraße 27, 1040 Vienna, Austria
[2] German Aerospace Center (DLR), Institute of Robotics and Mechatronics, Münchener Straße 20, 82234 Weßling, Germany
{zoltan.varga, esteve.valls.mascaro, daniel.sliwowski, dongheui.lee}@tuwien.ac.at

Abstract. Most research in deep learning focuses on a single modality, such as image, text, or proprioception data. However, humans benefit from leveraging information from diverse senses on a daily basis for richer information acquisition. Inspired by this, we design a transformer-based multimodal model for human action recognition and thoroughly evaluate its performance and robustness. Furthermore, we explore fusion methods to assess how modalities are best combined. Lastly, a model is trained to infer (generate) a missing modality. Our study shows that multimodal transformers perform better than their modality-specific equivalents. We achieve an improvement of 10.1% when using multiple data modalities over our vision-only baseline and outperform current state-of-the-art approaches by 32.8%. Furthermore, we evaluate a mean square error of 9.6% in the tactile force reconstruction task. The implemented model can be applied in scenarios where robotic assistance depends on recognising human actions for decision-making, tackling situations where vision is limited or audio and other modalities are required for deeper understanding.

Keywords: Artificial intelligence, Perception, Human action recognition, Machine Learning

1 Introduction

Despite the rapid growth of deep learning in the research of task understanding, most works focus on analysing the environment and actions purely based on visual perception [7, 21]. Limiting a model to visible knowledge fails to translate some of the embodied intelligence into robots. Approaching the topic from a biological point of view, humans have a total of five senses: sight, hearing, touch, taste, and smell. For various tasks in our daily lives, it is crucial to combine a handful of them and know which to focus on. Intelligent systems, such as robots,

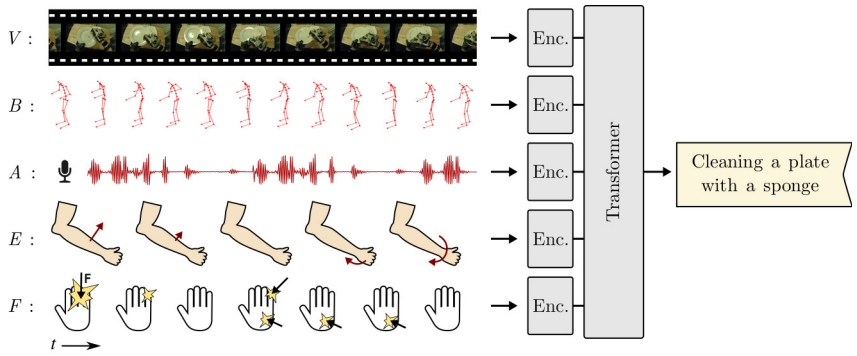


Fig. 1: Overview of a multimodal transformer model for human action classification. The visualised modalities are vision V, body skeleton pose B, audio A, muscle activity E from electromyography sensor (EMG), and tactile force F.

can emulate these senses by the means of sensors. Although the principles behind how the sensors work are different from how humans perceive the world, they fulfil the same role in understanding the environment and interacting with it.

For instance, through vision only, models often fail to distinguish between tasks such as grasping and squeezing, as the main difference lies in the applied gripping force; for example, opening a jar or a bottle requires adequate force, which cannot be determined only through visual means. To this end, we consider that leveraging multimodal data that extends from vision and text would improve the grounding of deep learning algorithms with reality.

Recognising human actions is a crucial part of human-robot interactions. To understand human actions, one needs to comprehend the evolution of the actions in time, allowing one to gain knowledge on temporally progressive tasks. For instance, by observing a human holding the fridge handle, one can not determine if the action is to close or to open the fridge. Understanding this sequence of motion primitives allows for excelling in the corresponding action recognition.

We consider the problem of recognising the action a human is performing in a kitchen scenario based on multiple sensory data, as shown in Figure 1. To this end, we propose a transformer-based model [23] to merge multiple modalities and study how the classification improves compared to that of single modalities. Additionally, we extend the model with a decoder that can infer one modality from the other (e.g., force from video data and muscle activity). This could be used for humanoid robots, allowing them to learn from demonstrations and mimic human movements.

To summarise, our contributions are as follows:

1. integration of five modalities with a transformer-based model for human action recognition;
2. thorough evaluation of different strategies for modality fusion;
3. exploration of regularisation techniques during training to enhance robustness in multimodal transformers.

2 Related Work

From a robotic point of view, human action recognition can be used in autonomous navigation systems [19], surveillance systems [3], and Human-Robot Interaction (HRI). Visual modalities, especially RGB videos, have been shown to be very effective for this purpose in deep learning [22]. Besides vision, many other modalities can be utilised for action recognition, such as skeleton pose [16] and audio [15].

Previous works for understanding actions (such as a Recursive Neural Network (RvNN) [13] and a Long Short-Term Memory (LSTM) [4, 12]) proposed to model the sequential dimension of the data by updating a representation that encapsulates the prior information. Although they can recognise temporal progress [4], they often perform poorly for complex patterns. Instead, we consider the use of a transformer [23] for this specific problem as it has shown outstanding performance for sequence modelling in different modalities like natural language [5], video prediction [2], and image classification [6, 24]. Transformer-based models aim to solve the limitations of prior approaches by using the attention mechanism to propagate information over time and learn the relationship between the short and long data sequences. Video Vision Transformers (ViViTs) [2] utilise the embedded spatiotemporal relations in videos. Our model is inspired by their approach of a factorised encoder [2].

While previous approaches only focused on single modalities [7, 15, 16], our work is motivated by the observation that leveraging information from different sources enriches the model's understanding of the actions. Recent studies have found different means of abstraction for multimodal problems. For instance, ImageBind [8] constructs a shared latent space for six different modalities, and as a result, it can perform cross-modal retrieval and zero-shot classification. Their model leverages modalities by pairing them with vision; meanwhile, our model fuses all of its modalities simultaneously. Multimodal neural networks differ in architecture regarding the exact way features of different modalities interact with each other. There is no consensus regarding the correct way of modality fusion, as a whole scale of various techniques is applicable. Some of the competing approaches are early fusion [5], late fusion [9], or concatenation [26]. This work assesses this by comparing different fusion strategies in multimodal learning.

Despite the main research being human action recognition, previous works [1, 14, 25] have considered the task of modality inference. Observing the same action from different modalities carries a redundancy of information. This can be used to infer missing data using other modalities, i.e., sensors that do not directly measure the data of interest. Inspired by other works [11, 23], we extend our model with a decoder to recreate a missing modality instead of classification. This experiment demonstrates our model's ability to understand the relationship between different modalities. In most prior works, Inertial Measurement Unit (IMU) data is inferred from vision [14, 18]. Our work differs from these models as it has multiple input modalities inferring a single output. Cross-modal inference

can appear in various robotic applications, including teleoperation [1], and it has been shown to improve object recognition as well [25].

3 Methodology

The goal of action recognition is to determine the action a human is performing a_i based on a trimmed video clip $V_i = \{V[t_i] \dots V[t_i + T]\}$ of length T. In our work, we consider additional data sources that are available and synchronised to the frames V_i in the video recording; namely, body pose B_i, tactile force data F_i, muscle activation E_i, and audio recording captured between frames t_i and $t_i + T$.

3.1 Model Architecture

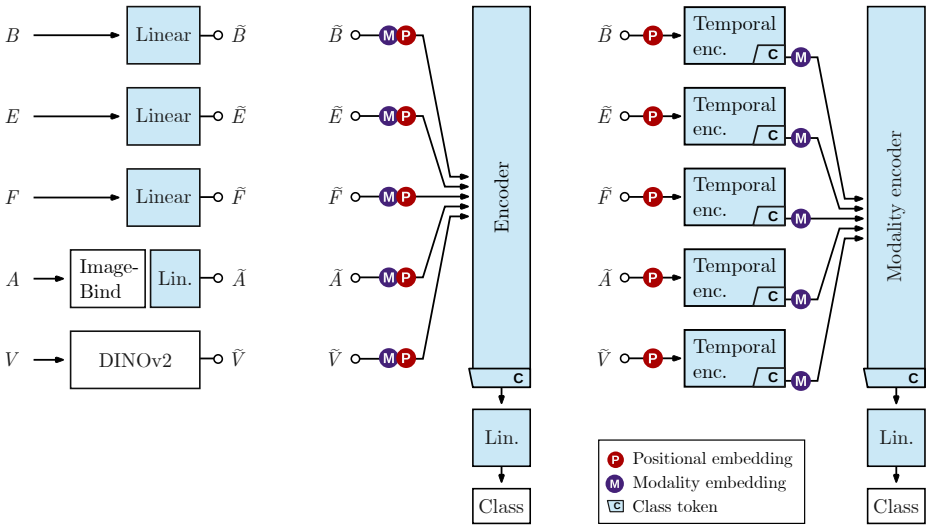
First, we require that information from all modalities be projected into an embedding space of the same size d. For most modalities, like body skeleton, tactile force, and EMG data, we use a linear layer to project the data. In the case of the vision modality, we use a frozen DINOv2 backbone [20] to compute the image features $\tilde{V} \in \mathbb{R}^d$ for each frame in the recording $V[t_i] \dots V[t_i + T]$. Finally, we compute the spectrograms of the audio signal and encode them with a frozen ImageBind [8] audio encoder, then project them via linear layer to create audio features $\tilde{A} \in \mathbb{R}^d$. We freeze pre-trained backbones for the visual and audio features as this speeds up the training process and makes the model less prone to overfitting.

Transformer Encoder. After creating tokens from raw data, they get passed to the encoder. To encode the temporal information about the data, we leverage a sinusoidal positional embedding and add it to each sample in each modality. Additionally, to allow the model to differentiate between the modalities, we learn a modality embedding M, which we add to each modality.

We employ a similar strategy for aggregating the information from the transformer output tokens as ViT [6]; namely, we prepend the tokens with a learnable class token and use its corresponding output token in further stages of the model.

Modality Fusion in the Encoder. We consider two different procedures for fusing tokens of separate modalities (Figure 2). On one hand, in *early fusion*, the temporal sequence of each modality gets stacked together before encoding. This fusion method has only one class token, summarising information along the time axis and among modalities. After adding the class token, the sequence consists of $m \cdot f_s \cdot T + 1$ tokens, where m denotes the number of modalities.

On the other hand, in *late fusion*, each modality has a separate encoder. The temporal encoders process information along the time axis, focusing on one modality at a time. Each of the m encoders works with $f_s \cdot T$ tokens and an additional class token that is not shared among modalities. After the first block



(a) Creating tokens from data. This is the trunk of all our models.

(b) Early fusion encoder

(c) Late fusion encoder

Fig. 2: Overview of the model architecture. Modules with trainable parameters are highlighted with a light blue background.

of encoders, only the class tokens are passed to the modality encoder. They are stacked into a sequence where modality embeddings and a new class token are added.

3.2 Implementation Details

We chose cross entropy as the loss function. The model is trained using the AdamW optimiser [17] in PyTorch. We use a cosine learning rate scheduler starting at $\eta_0 = 10^{-5}$. Our early fusion model has 7.45 million of learnable parameters, and the late fusion model has 42.9 million. The DINOv2 [20] and the ImageBind [8] modules have 22.1 and 85.4 million parameters, respectively, which remain frozen. All videos were sampled down to $f_s = 5$ frames per second to spare computation time and storage space. We divided actions into segments of length $T = 10$ s.

4 Experimental setup

We consider the ActionSense dataset [4] as it contains various sensory data like body and finger tracking, muscle activity, tactile, audio, and video data from different perspectives in a kitchen scenario where a person performs simple tasks. Food preparation and corresponding household chores are some of the most

desired topics in the field of automation and robotics. We have followed the preprocessing of [4] during the experiments to ensure a feasible baseline comparison with their LSTM model. The results of the LSTM have been recalculated here and do not match the results reported in [4].

The visual modality was provided by camera number 5 and the egocentric camera of [4]. We apply the same strategy for preprocessing the audio [8] and tactile data [4] as previous works. For the audio, we compute spectrograms using mel-scaled frequency banks [10], and for the tactile data, we use average pooling to reduce the size of the data from $\mathbb{R}^{16\times16}$ to $\mathbb{R}^{4\times4}$.

Following the standard evaluation for ActionSense [4], we adopt cross-validation among the five subjects of the dataset. The model with the lowest loss is selected to be the final model of the training session.

5 Evaluation

The evaluation compares uni- and multimodal models based on their understanding of human actions. We measure the action recognition ability of models by classification accuracy.

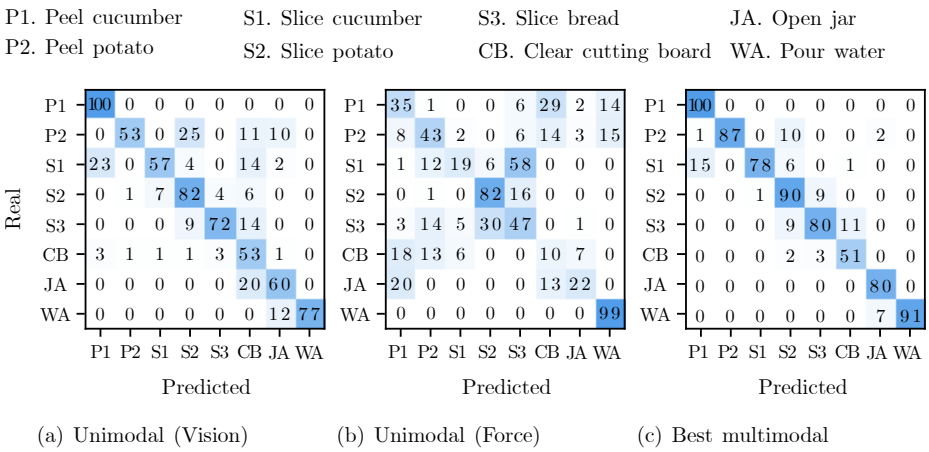
Table 1: Ablation study for uni- and multimodal models. The model that achieved the highest accuracy is highlighted in bold, and the second highest is underlined. Multimodal models performing worse than the highest corresponding unimodal model are marked with an arrow. All values depict accuracies in %.

Modality	Ours	LSTM
F. Force	33.5	20.8
E. EMG	22.4	17.1
B. Body	32.9	24.8
A. Audio	<u>37.9</u>	–
V. Video	**71.7**	–

Modalities					Mask	Ours		LSTM
F	*E*	*B*	*A*	*V*		Early f.	Late f.	
✓		✓				41.2	44.4	26.1
✓	✓	✓				41.9	47.6	↓ 24.5
✓				✓		<u>77.7</u>	77.5	–
✓	✓	✓	✓	✓		75.7	74.6	–
✓	✓	✓	✓	✓	$\mathcal{M}_R^{0.5}$	↓ 70.6	**81.8**	–
✓	✓	✓	✓	✓	$\mathcal{M}_V^{0.5}$	↓ 70.9	77.1	–

5.1 Unimodal Classification

In Table 1, we report the evaluation of our Transformer-based model compared to the LSTM baseline [4] for the human action classification from single modalities. The results showcased how vision is the most informative modality as it helps to distinguish object types (i.e. cucumber vs potato) and implicitly carries most of the other modalities: body pose can be inferred from vision, as well as the applied force [18]. Finally, our results showcase an improvement over each internal modality (force, EMG and body) compared to the baseline [4].



(a) Unimodal (Vision) (b) Unimodal (Force) (c) Best multimodal

Fig. 3: Showing the advantage of using multiple modalities by comparing selected parts of confusion matrices.

5.2 Multimodal Models

The multimodal model recognises many instances correctly that are misclassified by the vision-based model. As Figure 3 shows, using perception only often fails to identify objects and confuses *Peeling* with *Slicing* due to occlusions. Replacing the unimodal video model with a multimodal one increases the lowest precision from 39% to 52% and the lowest recall from 40% to 64%.

However, we observe that the redundancy existing across different modalities (e.g., observing a human cutting with sound, motion, and forces) causes the model to overfit during training. To cope with this issue, we freeze the pre-trained encoders (DINOv2 [20] and ImageBind [8]). Additionally, we propose a masking strategy to regularise the dependencies of the model to high-influential modalities, like vision. We provide more details in *Masking the Encoder*.

Attention Scores. The multimodal model with late fusion separates the time evolution from the modality fusion, allowing us to compare modalities without time dependency. We calculated the average attention score over the heads and layers of all attention blocks and extracted the row corresponding to the class token. Figure 4 shows that video modality alone gets more attention than all other tokens combined. Other modalities scored around 1%, with the tactile force being higher. The learnable parameter that serves as the initial class token is of little importance, as it does not hoard any information about the inputs.

We observe that for some human actions, such as *Peeling* or *Slicing*, the internal sensors gain influence in the decision-making of our transformer. We believe that, given the small manipulation area, the vision might be occluded, and the model relies on more fine-grained sensors like force and body pose to determine these actions. The video modality gets high attention for all classes

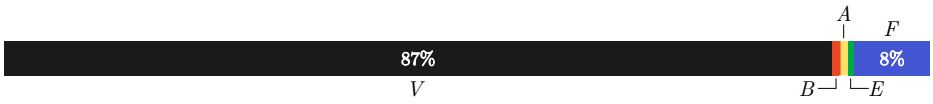


Fig. 4: Attention score of each modality in late fusion. (V: vision, B: body skeleton pose, A: audio, E: muscle activity, F: tactile force)

without exception but the highest for *Spreading something on bread*. That is because jelly and almond butter cannot be distinguished without vision. Colour is the only cue in that case, as the dataset lacks a taste sensor.

Masking the Encoder. During the experiments, we noticed a strong dependency of the multi-modal models on the vision. To overcome this, we add regularisation in the form of masking, similar to [11]. This process makes the model focus more on the remaining tokens and acts as a powerful form of regularisation if applied stochastically. Our implementation generates masks separately for each element of the batched input.

We evaluate the influence of masking the modalities on the performance of the model. We denote masking a modality X with probability p as $\mathcal{M}_X^p$. For $X = R$, we consider randomly masking any modality. Table 1 shows that by applying masking we can improve the performance of the late fusion models by an additional 7.27%. We achieve the highest performance (81.8%) with $\mathcal{M}_R^{0.5}$. We do not observe a performance increase in the early fusion model; we consider this to be caused by the high number of tokens the encoder has to process simultaneously which increases the complexity of the task and reduces overall performance.

As Table 1 shows, late and early fusion have similar accuracies without regularisation. By adding a training mask, the late fusion model overfits significantly less, achieving higher accuracies, while early fusion does not benefit from it. We believe that the complexity of the dependencies between the sequence of tokens is simpler in the late fusion, as we process each modality first and later combine it. On the contrary, in early fusion, both steps are done simultaneously, extending the context length of the transformer. By incorporating a modality-specific encoder in late fusion, our model achieves higher performance and showcases stronger benefits of leveraging multimodal data.

5.3 Robustness Study

In real scenarios, merging multiple modalities has additional benefits besides enhancing the action recognition capabilities. For instance, a person can identify water boiling despite being in a different room and can peel an apple while watching the phone. Thus, in this section, we showcase the benefit of multimodal models when vision is not given at inference time but still used during training. In a robotics scenario, this would be the case when the camera sensor malfunctions or is under bad lighting conditions. Testing for robustness shows whether

the model can still recognise actions after one of the sensors stops functioning. We aim to demonstrate that using multiple modalities has advantages besides enhancing the recognition performance.

Removing a sensor entirely undoubtedly causes a drop in accuracy, but at a lower extent for masked models. As shown in Figure 5, they achieve similar accuracy as the model that never had access to vision. The unmasked model strongly depends on the visual input and completely fails when excluded. To achieve a feasible comparison, we evaluate some models on the multimodal dataset without vision (equivalent to masking with $\mathcal{M}_V^1$). Then, we compare them to the experiment where vision was still available and examine the extent of the drop in accuracy.

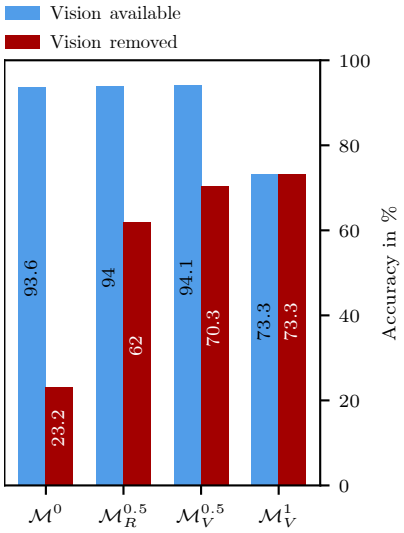


Fig. 5: Performance achieved by different masks in the training phase. This experiment uses the egocentric camera as the visual modality and trains the model for a simplified classification objective.

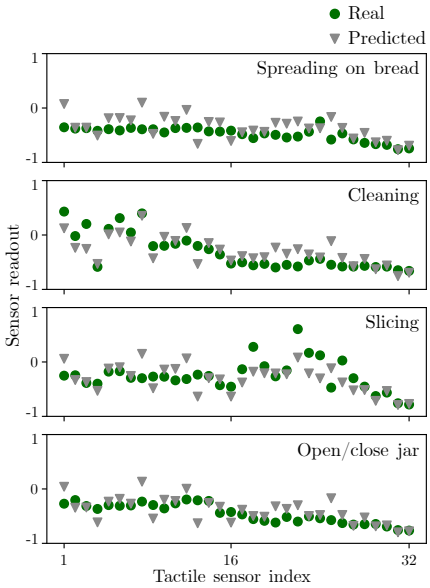


Fig. 6: Inferring tactile force F from other modalities (body skeleton B, muscle activity E, audio A and video V).

5.4 Modality Inference

We evaluate a model trained to infer a missing modality using features from other modalities. Although not part of the main human action recognition task, this experiment demonstrates our model's ability to understand the relationship between different modalities.

The force modality was removed from all data inputs to simulate the missing modality. Then, the model was trained to reconstruct the missing data using body skeleton, EMG, audio, and video. Figure 6 plots the predicted force sensor readouts against the ground truth. The force sensors with indices 1–16 are located on the left hand of the subject, and those with 17–32 on the right one. We are able to reconstruct the tactile force information with an average Mean Square Error (MSE) of 9.6%.

6 Conclusion

Most deep learning research for action recognition concentrates solely on vision, yet visual modalities alone provide insufficient information for many robotic applications. Aiming for efficient and robust interactions with a physical environment raises the need for combining multiple modalities. To this end, we introduce a multimodal transformer model capable of recognising kitchen tasks based on five different modalities (force, muscle activity, skeleton pose, audio, and video).

We evaluate our proposed design and motivation through an extensive ablation study, which showcases how intelligently combining modalities enhances the performance in human action classification while being more robust to potential sensor failures. In fact, our multimodal approach outperforms baselines as well as modality-specific models across all tasks and variations. Our results showcase that explicitly biasing the model to randomly drop the attention to certain modalities during training acted as a regulariser, which enhances both accuracy and robustness, leading to an 81.8% of accuracy, 10.1% higher than the modality-specific counterpart.

Additionally, we investigate the optimal method to fuse multiple modalities. Our results indicate that a late fusion approach enhances the classification compared to early fusion. This difference becomes more significant when regularisation techniques, such as masking, are applied.

In conclusion, our proposed multimodal transformer model benefits strongly from the use of different sensors for action classification and opens new possibilities for their application in robotics, as well as for further extension to enhance the task of robot imitation. We believe that larger multimodal datasets would be valuable for further research with manipulation tasks, and we plan to explore those in the future.

Acknowledgments

This work has been partially supported by the European Union project INVERSE under grant agreement No. 101136067.

Glossary

V, B, A, E, F	Input features of modalities vision, body skeleton pose, audio, muscle activity and tactile force, respectively
$\tilde{V}$, $\tilde{B}$, $\tilde{A}$, $\tilde{E}$, $\tilde{F}$	Feature tokens calculated by modality-specific backbones
$\mathcal{M}_X^p$	Masking modality X with probability p. For $X = R$ we consider a randomly selected modality

References

1. H. Ahn, Y. Michel, T. Eiband, and D. Lee. Vision-based approximate estimation of muscle activation patterns for tele-impedance. *IEEE Robotics and Automation Letters*, 2023.
2. A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid. Vivit: A video vision transformer. *ICCV*, 2021.
3. S. Danafar and N. Gheissari. Action recognition for surveillance applications using optic flow and svm. In *ACCV*, 2007.
4. J. DelPreto, C. Liu, Y. Luo, M. Foshey, Y. Li, A. Torralba, W. Matusik, and D. Rus. Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. *NeurIPS*, 2022.
5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 2018.
6. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition. *ICLR*, 2021.
7. C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spatiotemporal residual networks for video action recognition. *CoRR*, 2016.
8. R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
9. R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra. Omnivore: A single model for many visual modalities. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
10. Y. Gong, Y.-A. Chung, and J. Glass. Ast: Audio spectrogram transformer. *arXiv:2104.01778*, 2021.
11. K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF conference on computer vision and pattern recognition*, 2022.
12. S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 1997.
13. M. Javeed, N. A. Mudawi, B. I. Alabduallah, A. Jalal, and W. Kim. A multimodal iot-based locomotion classification system using features engineering and recursive neural network. *Sensors*, 2023.
14. Y. Li, Y. Du, C. Liu, F. Williams, M. Foshey, B. Eckart, J. Kautz, J. B. Tenenbaum, A. Torralba, and W. Matusik. Learning to jointly understand visual and tactile signals. In *ICLR*, 2023.
15. D. Liang and E. Thomaz. Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2019.

16. J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang. Skeleton-based action recognition using spatio-temporal lstm network with trust gates. *IEEE transactions on pattern analysis and machine intelligence*, 2017.
17. I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017.
18. N. Louis, J. J. Corso, T. N. Templin, T. D. Eliason, and D. P. Nicolella. Learning to estimate external forces of human motion in video. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.
19. M. Lu, Y. Hu, and X. Lu. Driver action recognition using deformable and dilated faster r-cnn with optimized region proposals. *Applied Intelligence*, 2020.
20. M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.
21. H. H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin. Video-based human action recognition using deep learning: A review, 2022.
22. Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, 2022.
23. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017.
24. R. Winastwan. Image classification with vision transformer. *Towards Data Science*, 2023.
25. Y. Zhang, X. Ding, K. Gong, Y. Ge, Y. Shan, and X. Yue. Multimodal pathway: Improve transformers with irrelevant data from other modalities. *arXiv:2401.14405*, 2024.
26. Y. Zhang, K. Gong, K. Zhang, H. Li, Y. Qiao, W. Ouyang, and X. Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv:2307.10802*, 2023.