Supplementary information

Skew-pair fusion theory: An interpretable multimodal fusion framework

Contents

1 Supplementary Tables 2
1.1 Text-Audio Fusion Experiments 2
1.2 Audio-Video Fusion Experiments 3
1.3 Image-Text Fusion Experiments 4
2 Supplementary Figures 6
2.1 The Concept of Multimodal Fusion 6
2.2 Late-Fusion Strategy 7
2.3 Pairwise-Fusion Strategy 8
2.4 Clustering Analysis using *t*-SNE 8
3 Datasets 10
References 10

1 Supplementary Tables

1.1 Text-Audio Fusion Experiments

We conducted a comprehensive analysis of text-audio fusion using BERT[1] and Wav2Vec2[2] models on the IEMOCAP benchmark dataset for multimodal emotion recognition[3].

To elucidate the contribution of individual layers to task performance, we evaluated each layer of BERT and Wav2Vec2 in isolation. Our results revealed two key trends: (i) the performance of both models does not consistently improve with increasing layer depth, suggesting a non-monotonic relationship between layer depth and task performance; and (ii) the performance trends across layers diverge between the two models, underscoring the importance of considering cross-modal heterogeneity in multilayer models. These findings highlight the complex, nonlinear nature of hierarchical representation learning in multilayer models and underscore the need for careful consideration of layer-wise contributions in multimodal fusion tasks.

Supplementary Table S1 | **Single-layer characterization performance** (Evaluation Index: Accuracy (*ACC*))

Layer	BERT (%)	Wav2Vec2 (%)
12	**69.8**	**68.3**
11	69.1	67.6
10	67.4	66.7
9	65.1	65.0
8	61.2	62.4
7	62.0	59.4
6	61.8	54.0
5	61.2	51.0
4	60.2	48.3
3	58.7	45.8
2	52.5	46.1
1	44.0	43.4

In single-layer fusion experiments, we investigated the impact of fusion direction on task performance by conducting two complementary sets of experiments: (i) fusing the final layer of BERT with each layer of Wav2Vec2, and (ii) reversing the fusion direction as a control, where each layer of Wav2Vec2 was fused with the final layer of BERT. The results reveal a striking asymmetry in the fusion paradigm, with the optimal fusion pair differing significantly depending on the direction of fusion. Notably, these findings suggest that the best-performing layers in unimodal experiments do not necessarily form the optimal fusion pair in the multimodal setting. Instead, our results highlight the importance of carefully curating fusion pairs to account for the inherent heterogeneity between individual modalities. This heterogeneity underscores the need for a more nuanced approach to multimodal fusion, one that considers the complex interactions and dependencies between different modalities.

Supplementary Table S2 | **Fusion heterogeneity evaluation** (Evaluation Metric: Unweighted Average (*UA*))

BERT Layer	Wav2Vec2 Layer	UA

12	12	76.06%
	11	75.57%
	10	75.05%
	9	76.47%
	8	**76.57%**
	7	74.45%
	6	75.26%
	5	74.53%
	4	74.84%
	3	73.54%
	2	73.54%
	1	72.54%
Wav2Vec2 Layer	**BERT Layer**	**UA**
12	12	**76.06%**
	11	75.27%
	10	74.85%
	9	70.47%
	8	68.77%
	7	68.37%
	6	67.45%
	5	67.26%
	4	67.53%
	3	67.05%
	2	67.84%
	1	67.39%

1.2 Audio-Video Fusion Experiments

We conducted a comprehensive analysis of audio-video fusion using Wav2Vec2[2] and TimeSformer[4] models on the AVE (Audio Visual Event Localization) dataset in the temporal localization tasks[5].

To elucidate the contribution of individual layers to task performance, we evaluated the performance of each layer in Wav2Vec2 and TimeSformer separately (**Supplementary Table S3**). Our results reinforce the paradigm of dual cross-modal heterogeneity observed in our previous text-audio fusion experiments (**Supplementary Table S1**), highlighting the complexity of multimodal representation learning. Notably, we found that the optimal performance for Wav2Vec2 and TimeSformer was achieved in the 8th and 9th layers, respectively, rather than the final layers. This finding challenges the prevailing assumption that the final layers of cross-modal models are optimal for fusion-pair alignment, suggesting that conventional late-fusion strategies may be based on a flawed hypothesis. Specifically, our results indicate that the coupling of the last abstract layers of cross-modal models is not necessarily the most effective approach for fusion-pair alignment, underscoring the need for a more nuanced understanding of multimodal representation learning and the development of more effective fusion strategies.

Supplementary Table S3 | Single-layer characterization performance (Evaluation Index: Accuracy (*ACC*))

Layer	Wav2Vec2 (%)	TimeSformer (%)
12	67.7	71.9

11	66.7	71.1
10	66.4	69.4
9	67.4	**72.9**
8	**69.1**	66.2
7	68.1	61.7
6	66.7	61.0
5	65.7	55.0
4	63.9	46.2
3	64.2	39.1
2	60.7	30.1
1	49.3	16.5

In single-layer fusion experiments, we investigated the impact of fusion direction on task performance by conducting two complementary sets of experiments: (i) fusing the final layer of TimeSformer with each layer of Wav2Vec2, and (ii) reversing the fusion direction as a control, where each layer of Wav2Vec2 was fused with the final layer of TimeSformer (**Supplementary Table S4**). Our results further reinforce the paradigm of dual cross-modal heterogeneity observed in our previous text-audio fusion experiments (**Supplementary Table S2**), demonstrating that the optimal fusion pair is highly dependent on the direction of fusion and the specific layers involved. This finding underscores the complexity of multimodal representation learning and highlights the need for a more nuanced understanding of the interactions between different modalities.

Supplementary Table S4 | Fusion heterogeneity evaluation (Evaluation Index: Accuracy (*ACC*) and Unweighted Average (*UA*))

TimeSformer Layer	Wav2Vec2 Layer	Accuracy
12	12	78.86%
	11	79.10%
	10	79.85%
	9	81.59%
	8	80.60%
	7	79.85%
	6	81.34%
	5	79.85%
	4	**81.84%**
	3	78.61%
W2V2 Layer	**TimeSformer Layer**	**UA**
12	12	78.86%
	11	**81.10%**
	10	79.85%
	9	79.60%
	8	79.10%
	7	79.60%
	6	76.87%
	5	74.88%
	4	75.37%
	3	71.39%

1.3 Image-Text Fusion Experiments

We conducted a comprehensive analysis of multimodal hate speech detection using a text-based BERT model[6] and an image-based BEiT model[7], both specifically fine-tuned for the Hateful

Memes dataset. Following established fine-tuning practices, we optimized the models' performance by carefully adjusting their hyperparameters.

In single-layer experiments, we evaluated the performance of individual layers in BERT and BEiT separately (**Supplementary Table S5**). Our results reinforce the paradigm of dual cross-modal heterogeneity observed in our previous text-audio (**Supplementary Table S1**) and audio-video fusion (**Supplementary Table S3**) experiments, highlighting the complexity of multimodal representation learning. Notably, we found that the optimal performance of BERT and BEiT was achieved in the 12th (65.7%) and 10th layer (52.4%), respectively, rather than in symmetrical layers. This finding challenges the prevailing assumption that the coupling of symmetrical layers is optimal for fusion-pair alignment, suggesting that conventional pairwise-fusion strategies may be based on a flawed hypothesis. Instead, our results indicate that the optimal fusion pair may involve layers with different depths and abstraction levels, underscoring the need for a more nuanced understanding of multimodal representation learning and the development of more effective fusion strategies.

Supplementary Table S5 | **Single-Layer characterization performance** (Evaluation Index: *ROC AUC*)

Layer	BERT (%)	BEiT (%)
12	**65.7**	51.3
11	65.4	51.6
10	64.9	**52.4**
9	64.7	51.6
8	64.4	47.8
7	64.4	51.3
6	64.6	50.0
5	64.4	51.5
4	64.4	50.1
3	63.8	50.8
2	61.4	51.6
1	58.3	49.9

In our single-layer fusion experiments, we conduct two sets of experiments: (i) fusing the last layer of BERT with each BEiT layer, and (ii) reversing the fusion direction as a control. The results reveal a significant asymmetry between BERT and BEiT (**Supplementary Table S6**), aligning with previous text-audio (**Supplementary Table S2**) and audio-video fusion experiments (**Supplementary Table S4**). Specifically, our experiments demonstrate that the optimal fusion configuration (66.8%) involves coupling the last layer of BERT with the 10th layer of BEiT. This combination effectively leverages the high-level semantic information from the text modality and the mid-level visual information from the image modality. Conversely, reverse-staggered layer combinations generally lead to degraded performance.

Supplementary Table S6 | **Fusion heterogeneity evaluation** (Evaluation Index: *ROC AUC*)

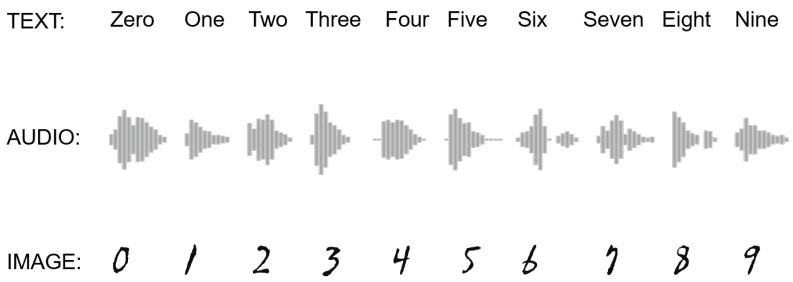
BERT Layer	BEiT Layer	ROC AUC
12	12	65.0%
	11	65.4%

	10	**66.8%**
	9	65.7%
	8	64.0%
	7	65.2%
	6	65.8%
	5	65.3%
	4	64.2%
	3	64.8%
	2	64.7%
	1	64.4%
BEiT Layer	**BERT Layer**	**ROC AUC**
12	12	**65.0%**
	11	64.9%
	10	62.2%
	9	62.5%
	8	62.4%
	7	62.1%
	6	62.7%
	5	62.4%
	4	62.3%
	3	61.8%
	2	59.4%
	1	57.3%

2 Supplementary Figures

2.1 The Concept of Multimodal Fusion

The concept of multimodal fusion, although still evolving and lacking consensus within the scientific community, has shown significant potential as a performance enhancer in various applications[8,9]. Conventionally, different modalities are represented as arrays of varying dimensions, such as 1D for sequences and signals, 2D for audio spectrograms or images, and 3D for video or volumetric images[10]. By leveraging the strengths of multiple modalities (**Supplementary Fig. S1**), multimodal fusion has opened new avenues for tackling complex tasks, including visual speech recognition, emotion recognition, video classification, object detection, and many other domains such as healthcare and biology.



Supplementary Fig. S1 | Modality representation of numbers 0-9. The text modality is rich in pragmatic and semantic information, whereas the audio modality captures phonological and acoustic features. In contrast, the image modality, specifically handwriting, provides visual and morphological cues.

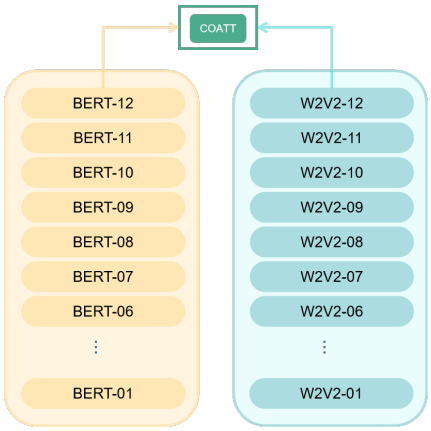
Despite the diversity of features across these modalities, multimodal neural networks do not always guarantee superior performance to their unimodal counterparts in downstream tasks. This counterintuitive observation can be attributed to the rule of inverse effectiveness in multisensory integration, which suggests that the integration of cross-modal stimuli does not necessarily lead to enhanced performance.

Remarkably, our text-text fusion task reveals that even within a seemingly "unimodal" fusion setting, the inherent heterogeneity between corpora with nuanced differences can be tantamount to a cross-modal scenario (**Figures 8** and **9**). Our results provide strong evidence for the universality of the skew-layer theory, which is substantiated by the rule of multisensory enhancement[11]. These findings challenge the conventional concept of multimodal fusion, which relies on integrating multiple data modalities via a shared forward model of neuronal activity[12,13,14]. Instead, our results suggest that skew-pair fusion offers substantial advantages over late- and pairwise-fusions in multifaceted domains.

However, a theoretical ensemble framework for aligning heterogeneous channels in latent representation spaces remains underdeveloped. Further research is needed to develop a comprehensive framework that can effectively integrate multiple modalities and models, and provide a deeper understanding of the underlying mechanisms of multimodal fusion[9,15,16].

2.2 Late-Fusion Strategy

The late-fusion approach involves processing each modality independently using a dedicated neural network (**Supplementary Fig. S2**). This strategy offers simplicity and minimizes the need for modality synchronization, making it a straightforward approach to implement. Late-fusion is particularly effective when each modality contains sufficient information to make robust decisions independently.

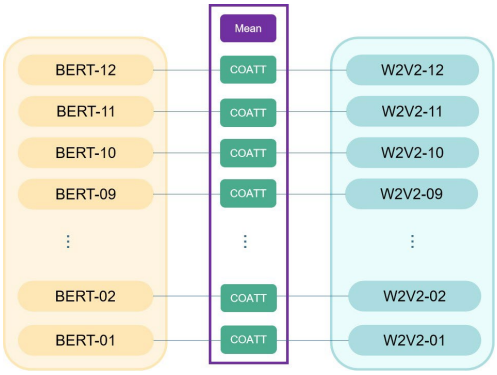


Supplementary Fig. S2 | Schematic diagram of late-fusion strategy. The late-fusion strategy involves combining the final layers of two neural network architectures: BERT (text-based) and Wav2Vec2 (audio-based). Each model processes its respective input stream through multiple layers, capturing abstract information at distinct levels. The coupling of the BERT-L12 and W2V2-L12 layers represents a candidate Fusion Pair (FP). The COATT module and Mean operation are employed to facilitate the fusion process.

However, late-fusion may not be optimal when interactions between individual modalities are critical for understanding the context. For instance, in emotion recognition tasks, the complementarity between audio and text may require an early-fusion approach to fully leverage their combined information. In such cases, late-fusion may not be able to capture the complex relationships between modalities, leading to suboptimal performance.

2.3 Pairwise-Fusion Strategy

The pairwise-fusion approach involves a hierarchical fusion process, where corresponding layers from the two neural networks are fused independently (**Supplementary Fig. S3**). This strategy enables the model to capture complex interactions between modalities at each level, resulting in a more nuanced and effective fusion process.



Supplementary Fig. S3 | Schematic diagram of pairwise-fusion strategy. The pairwise-fusion strategy involves symmetrically coupling corresponding layers from two neural network architectures: text-based BERT (from BERT-L01 to BERT-L12) and audio-based Wav2Vec2 (from W2V2-L01 to W2V2-L12). The COATT module and Mean operation are employed to facilitate the fusion process.

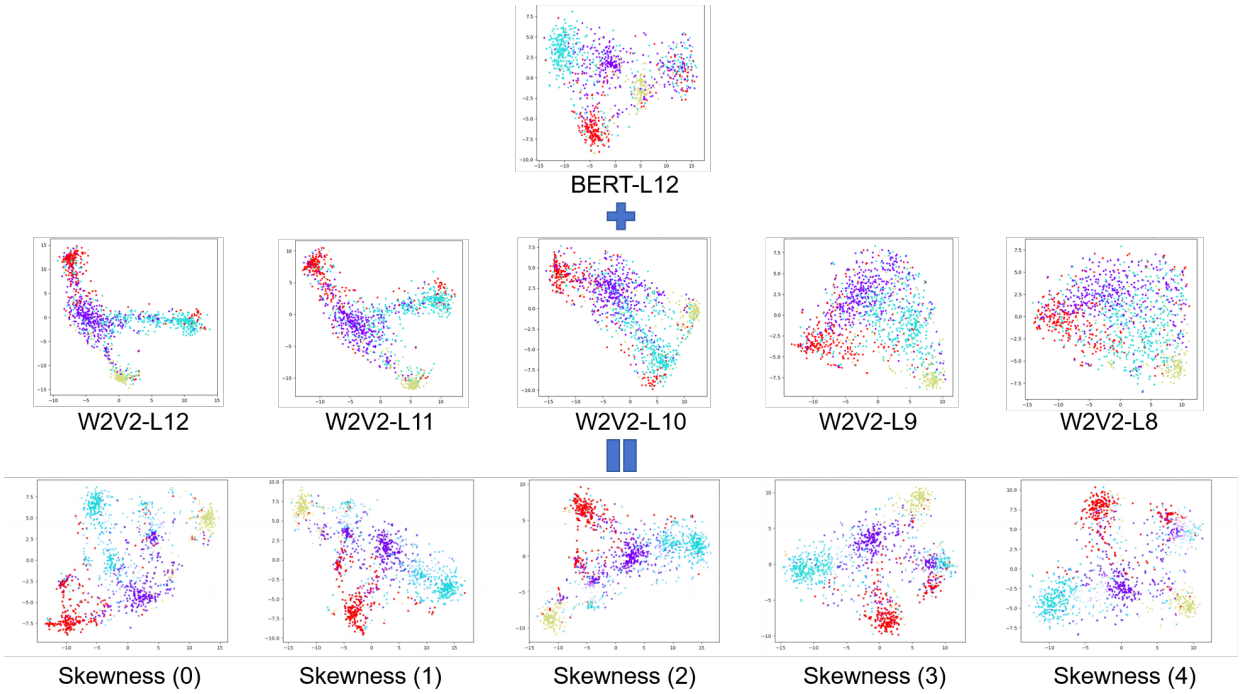
Pairwise-fusion offers several advantages, including the ability to precisely weight and integrate features at each layer. This facilitates a comprehensive understanding of the relationships between modalities, allowing the model to effectively leverage the strengths of each modality.

However, pairwise-fusion requires a certain degree of structural equivalence between individual modalities to ensure effective pairing and fusion. This can be a limitation, as it may not be possible to achieve perfect structural equivalence between modalities with different characteristics.

2.4 Clustering Analysis using *t*-SNE

To demonstrate the effectiveness of the skew-pair fusion strategy, we employed *t*-distributed stochastic neighbour embedding (*t*-SNE) clustering analysis[17] for both text-audio and text-text fusion experiments.

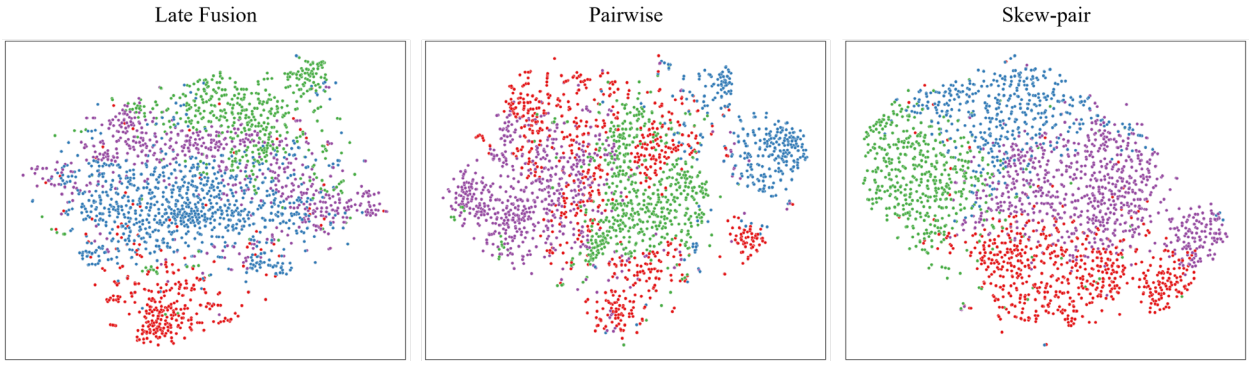
***t*-SNE Analysis of Text-Audio Fusion**. We conducted a t-SNE clustering analysis using the BERT and Wav2Vec2 models on the IEMOCAP dataset (**Supplementary Fig. S4**). By applying *t*-SNE to reduce dimensionality, we visualized the competitive performance of varying skewness ($d \in [0, 4]$), which informed the fine-tuning of the hyperparameter skewness (d). Our analysis revealed distinct clustering patterns, indicating that the models encode relatively independent information in their representation spaces.



Supplementary Fig. S4 | Visualizing clustering patterns with *t*-SNE. In the single-layer Text-Audio Fusion Experiment, we utilize BERT and Wav2Vec2 models on the IEMOCAP benchmark dataset, employing skew-pair fusion strategy for multimodal emotion recognition. To illustrate the clustering patterns, we apply *t*-SNE to reduce dimensionality and visualize the competitive performance of varying skewness ($d \in [0, 4]$).

The difference in clustering patterns can be attributed to the distinct characteristics of different modality data during model training. BERT focuses on semantic information in text, while Wav2Vec2 emphasizes acoustic properties in audio. Notably, the similarity in patterns between the final layer of BERT and the 8th layer of Wav2Vec2 suggests shared attributes or complementary information in the representation space. This implies that information fusion between these layers may enhance the model's expressive power while maintaining their respective modality characteristics.

***t*-SNE Analysis of Text-Text Fusion**. To further demonstrate the strength of the skew-pair fusion strategy, we compared its performance to late-fusion, pairwise-fusion, and skew-pair fusion strategies in the previous Text-Text Fusion Experiment using *t*-SNE (**Supplementary Fig. S5**).



Supplementary Fig. S5 | Clustering patterns of fusion strategies. This figure illustrates the clustering patterns of late-fusion, pairwise-fusion, and skew-pair fusion strategies using *t*-SNE in the topic classification task. Each coloured point represents a sample with a topic label, and different colours indicate distinct topic classifications. The visualization demonstrates the superior clustering effect of the skew-pair fusion strategy compared to the other two.

The results clearly demonstrate the superior clustering effect of the skew-pair fusion strategy strategy ($d = 0,\ \eta = 1/12$) compared to the other two approaches. This highlights the ability of skew-pair fusion to effectively integrate information from individual modalities, resulting in a more cohesive and meaningful representation of the data.

3 Datasets

The IEMOCAP dataset[3] consists of 151 videos of recorded dialogues, featuring 2 speakers per session, totalling 302 videos. Each segment is annotated for 9 emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed, and neutral) as well as valence, arousal, and dominance. The dataset spans 5 sessions with 5 pairs of speakers.

The AVE (Audio-Visual Event Localization) dataset[5] is a comprehensive collection of over 4000 videos, spanning 28 event categories. Each video is annotated to indicate the temporal boundaries of audio-visual events. The dataset encompasses a diverse array of events, including human activities, animal behaviors, music performances, and vehicle sounds. Each category has at least 60 videos, with some containing up to 188 videos. We investigate three temporal localization tasks: supervised and weakly-supervised audio-visual event localization, and cross-modality localization. We utilize audio-based Wav2Vec2 and video-based TimeSformer models for these tasks[5].

The Hateful Memes dataset[6] is a multimodal dataset for hateful meme detection (image + text) containing over 10,000 examples created by Facebook AI. Images were licensed from Getty Images to support researchers in their work.

The C4EL is a novel parallel corpus comprising two semantically equivalent datasets: a Latin-English code-mixed corpus (C4EL.LE) and its monolingual English counterpart (C4EL.EN). The metadata (C4.EN) includes text, timestamp, and URL for each entry.

References

1. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. in *Proceedings of the 2019 Conference of the North American* 4171–4186 (Association for Computational Linguistics, 2019). doi:10.18653/v1/N19-1423.
2. Baevski, A., Zhou, H., Mohamed, A. & Auli, M. Wav2Vec 2.0: A framework for self-supervised learning of speech representations. in *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)* 1–12 (Neural Information Processing Systems (NeurIPS), 2020).
3. Busso, C. *et al.* IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **42**, 335–359 (2008).
4. Bertasius, G., Wang, H. & Torresani, L. Is space-time attention all you need for video understanding? in *Proceedings of the 38th International Conference on Machine Learning Machine Learning* 813–824 (2021).
5. Tian, Y., Shi, J., Li, B., Duan, Z. & Xu, C. Audio-Visual event localization in unconstrained videos. in *Computer Vision - ECCV 2018* (eds. Ferrari, V., Hebert, M., Sminchisescu, C. & Weiss, Y.) 252–268 (Springer Nature, 2018). doi:10.1007/978-3-030-01216-8_16.
6. Kiela, D. *et al.* The hateful memes challenge: Detecting hate speech in multimodal memes. in *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing System* 2611–2624 (ACM, 2020).
7. Bao, H., Dong, L., Piao, S. & Wei, F. BEiT: BERT pre-training of image transformers. (2021).
8. Baltrusaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 423–443 (2019).
9. Fei, N. *et al.* Towards artificial general intelligence via a multimodal foundation model. *Nat. Commun.* **13**, 3094 (2022).
10. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
11. Stein, B. E. & Stanford, T. R. Multisensory integration: Current issues from the perspective of the single neuron. *Nat. Rev. Neurosci.* **9**, 255–266 (2008).
12. Meredith, M. A. & Allman, B. L. Subthreshold multisensory processing in cat auditory cortex. *Neuroreport* **20**, 126–131 (2009).
13. Budinger, E., Heil, P., Hess, A. & Scheich, H. Multisensory processing via early cortical stages: Connections of the primary auditory cortical field with other sensory systems. *Neuroscience* **143**, 1065–1083 (2006).
14. Friston, K. J. Modalities, modes, and models in functional neuroimaging. *Science* **326**, 399–403 (2009).
15. Osorio, D. Interpretable multi-modal data integration. *Nat. Comput. Sci.* **2**, 8–9 (2022).
16. Pahud de Mortanges, A. *et al.* Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. *npj Digit. Med.* **7**, 195 (2024).
17. Maaten, L. van der & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).