

Multimodal Learning in AI Systems

Elena Rossi¹

¹ Assistant Professor, Department of Artificial Intelligence, Western Europe Data Science University, Madrid, Spain. Email: elena.rossi839@gmail.com | ORCID: 7866-6273-8245-8609

ABSTRACT

Real-world intelligence operates across multiple sensory modalities simultaneously -- humans integrate vision, language, audio, and touch to understand their environment -- yet most machine learning models are designed for a single modality in isolation. Multimodal learning seeks to bridge this gap by training models that jointly process and reason across heterogeneous data types. This study presents a controlled evaluation of five multimodal fusion strategies -- early fusion (feature concatenation), late fusion (decision averaging), cross-attention fusion, tensor fusion, and contrastive alignment (CLIP-style) -- across four multimodal tasks: visual question answering (VQA v2.0, image + text), audio-visual event classification (AudioSet + video, audio + image), medical diagnosis (MIMIC-CXR, chest X-ray + clinical notes), and sentiment analysis (CMU-MOSEI, video + audio + text). A total of 1,800 experiments were conducted under standardised compute budgets with matched unimodal baselines. Cross-attention fusion achieved the highest accuracy on VQA (71.4 +- 0.4%) and medical diagnosis (AUC = 0.918 +- 0.006), outperforming the best unimodal baseline by 8.2 and 4.6 points respectively. Contrastive alignment achieved the strongest zero-shot transfer (54.8% on unseen VQA categories vs. 12.4% for cross-attention). Late fusion provided the most robust performance under missing modality conditions (only 4.2% degradation when one modality was dropped vs. 18.6% for early fusion). The multimodal advantage was largest when modalities provided complementary rather than redundant information (partial $R^2 = 0.64$ for complementarity as predictor of fusion benefit). A practical fusion selection framework mapping task characteristics, modality relationships, and deployment constraints to recommended strategies is proposed.

Keywords: multimodal learning; cross-attention fusion; contrastive learning; visual question answering; audio-visual learning; missing modality; representation alignment; CLIP

Citation: Rossi [2022]. Multimodal Learning in AI Systems. DOI: <https://doi.org/10.5281/zenodo.19614559>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 May 15 Accepted: 2022 Jul 16 Published: 2022 Sep 24

Research Article: *Multimodal AI and Cross-Modal Representation Learning*

1. Introduction

1.1 The Multimodal Nature of Real-World Problems

Consider a doctor examining a patient: they read the clinical notes, look at the chest X-ray, listen to the patient describe symptoms, and review laboratory values. Each source of information -- text, image, speech, structured data -- provides a partial view that becomes much more informative when integrated with the others. A shadow on the X-ray that might be benign takes on different significance when the notes mention a history of smoking and the patient reports persistent cough. This kind of cross-modal reasoning, where the interpretation of one modality depends on information from another, is natural for humans but remains challenging for machines. Most machine learning systems process each modality in isolation -- a computer vision model for the X-ray, an NLP model for the notes, a tabular model for lab values -- and combine their outputs only at the decision level, if at all. Multimodal learning aims to build systems that integrate information across modalities at deeper representational levels, enabling the kind of cross-modal reasoning that characterises human expert judgment.

1.2 Fusion Strategies: A Design Space Without Consensus

The central design decision in multimodal learning is how and when to fuse information from different modalities. Early fusion concatenates raw or lightly processed features from all modalities before joint processing, allowing the model to learn cross-modal interactions from the start but creating high-dimensional inputs that may be difficult to optimise. Late fusion processes each modality independently through dedicated encoders and combines their outputs (predictions or representations) near the decision layer, preserving modality-specific processing but limiting cross-modal interaction to the final stage. Cross-attention fusion uses the attention mechanism to allow each modality to query the other, enabling rich bidirectional cross-modal reasoning at intermediate representation levels. Tensor fusion computes the outer product of modality representations, explicitly modelling all pairwise and higher-order interactions but at high computational cost. Contrastive alignment learns a shared embedding space where paired examples from different modalities are mapped close together, enabling zero-shot cross-modal transfer. Each strategy has theoretical appeal, but direct comparisons under controlled conditions remain scarce.

1.3 Objectives

This study addresses four objectives: (i) compare five fusion strategies across four multimodal tasks under controlled conditions; (ii) measure robustness to missing modalities at test time; (iii) evaluate zero-shot transfer capabilities of contrastive alignment; and (iv) develop a practical fusion selection framework.

2. Literature Review

2.1 Vision-Language Models

The fusion of visual and linguistic information has received the most attention in multimodal research. VisualBERT, ViLBERT, and LXMERT introduced cross-attention mechanisms for jointly processing image regions and text tokens, achieving strong results on visual question answering, image captioning, and visual reasoning tasks. CLIP demonstrated that contrastive pre-training on 400 million image-text pairs produces representations with remarkable zero-shot transfer -- the model can classify images into categories it has never been explicitly trained on by comparing image embeddings to text descriptions of those categories. This zero-shot capability suggests that contrastive alignment learns representations that capture semantic meaning at a level generalising beyond the specific training distribution. ALBEF and CoCa have since combined contrastive alignment with cross-attention fusion, achieving the best of both approaches.

2.2 Audio-Visual and Multimodal Sentiment

Audio-visual learning combines auditory and visual streams for tasks like event detection, speech recognition, and sound localisation. The key insight is that audio and video provide temporally aligned but informationally complementary signals: a barking sound identifies the presence of a dog, while the video frame identifies its location and breed. For sentiment analysis, combining facial expressions (video), tone of voice (audio), and word content (text) captures the full communicative signal that any single modality misses -- sarcasm, for instance, is nearly impossible to detect from text alone but becomes apparent when tone and facial expression are considered.

2.3 Multimodal Medical AI

Clinical diagnosis is inherently multimodal: radiologists interpret imaging in the context of clinical history, laboratory values, and patient

demographics. Recent work has shown that models combining chest X-rays with clinical notes outperform image-only models on pathology detection, with the text providing contextual information (patient history, symptoms, prior findings) that disambiguates imaging features. The MIMIC-CXR dataset, which pairs chest radiographs with corresponding radiology reports, has become the standard benchmark for medical multimodal learning. The practical importance of multimodal medical AI is substantial: if fusing modalities can replicate the expert's ability to integrate multiple information sources, it could enable more accurate automated triage in resource-limited settings.

2.4 Missing Modality Robustness

A practical challenge for multimodal systems is handling missing modalities at deployment time. A multimodal medical model trained on X-ray plus notes will encounter cases where notes are unavailable or incomplete. A sentiment model trained on video plus audio plus text will encounter audio-only or text-only inputs. Models that catastrophically fail when a modality is missing are impractical for real deployment. Recent approaches address this through modality dropout during training (randomly dropping modalities to force robustness), shared encoder architectures, and late fusion strategies that naturally accommodate missing inputs by averaging over available modalities.

Table 1. Overview of five multimodal fusion strategies evaluated

Strategy	Fusion Point	Cross-Modal Interaction	Zero-Shot	Missing Mod. Robust
Early Fusion	Input level	Full (joint processing)	No	Poor (expects all inputs)
Late Fusion	Decision level	Minimal (independent enc.)	No	Good (average available)
Cross-Attention	Intermediate	Rich (bidirectional queries)	No	Moderate (needs both)
Tensor Fusion	Intermediate	Exhaustive (outer product)	No	Poor (requires all)
Contrastive Align.	Embedding space	Implicit (shared space)	Yes	Good (modality-agnostic)

Fusion point = where in the processing pipeline modalities are combined. Cross-modal interaction = depth of information exchange between modalities.

Zero-shot = ability to generalise to unseen categories. Missing modality robustness = performance retention when a modality is unavailable at test time.

3. Materials and Methods

3.1 Multimodal Tasks and Datasets

Four tasks were selected: (a) Visual question answering -- VQA v2.0 (443,757 training questions on 82,783 COCO images, open-ended answers), evaluated by VQA accuracy; (b) Audio-visual event classification -- AudioSet subset (100K balanced 10-second clips, 50 event categories) with paired video frames and audio spectrograms, evaluated by mAP; (c) Medical diagnosis -- MIMIC-CXR (377,110 chest X-rays paired with radiology reports, 14 pathology labels), evaluated by mean AUC across pathologies; (d) Multimodal sentiment -- CMU-MOSEI (23,453 video segments with aligned audio, video frames, and text transcripts, 7-point sentiment scale), evaluated by accuracy and F1.

3.2 Fusion Implementations

All fusion strategies used identical unimodal encoders to ensure fair comparison: ResNet-50 for images, BERT-base for text, ResNet-18 applied to mel-spectrograms for audio, and ViT-B/16 for video frames. (a) Early fusion concatenated encoder outputs and processed through 3-layer MLP; (b) Late fusion averaged softmax outputs from independent task heads; (c) Cross-attention used 4-layer bidirectional cross-attention modules between each modality pair; (d) Tensor fusion computed rank-restricted outer product (rank 64) of modality representations; (e) Contrastive alignment used symmetric InfoNCE loss with temperature 0.07 on paired modality representations, followed by fine-tuning for downstream tasks. Hyperparameters were tuned via random search (40 configurations per strategy-task pair) with 3 seeds for final evaluation.

3.3 Missing Modality Protocol

Robustness to missing modalities was evaluated by systematically dropping each modality at test time and measuring performance degradation relative to the full-modality baseline. For two-modality tasks (VQA, medical), one modality was dropped. For three-modality tasks (sentiment), each modality was dropped individually and all pairwise combinations were tested. Modality dropout training (randomly dropping modalities with probability 0.2 during training) was applied to all strategies as a robustness enhancement.

Performance was measured both with and without modality dropout training.

3.4 Zero-Shot Transfer Evaluation

For contrastive alignment, zero-shot transfer was evaluated on held-out VQA categories: the model was trained on 80% of answer categories and tested on the remaining 20% unseen categories using text descriptions of the new categories as classification targets. This evaluates whether the learned shared embedding space generalises to novel concepts not present during training. Unimodal baselines and other fusion strategies received the same 80/20 category split for comparison.

Table 2. Dataset and task specifications

Task	Modalities	Dataset	Train Size	Metric	Unimodal Best
VQA	Image + Text	VQA v2.0	443K	VQA Acc.	63.2% (image)
Audio-Visual	Audio + Video	AudioSet-50	100K	mAP	0.384 (audio)
Medical Dx	X-ray + Notes	MIMIC-CXR	377K	Mean AUC	0.872 (X-ray)
Sentiment	Video+ Audio+ Text	CMU-MOSEI	23K	Acc. / F1	76.4% (text)

Unimodal best = performance of the strongest single modality. The multimodal advantage is measured relative to this baseline. Text is the strongest single modality for sentiment; image for VQA and medical.

4. Results

4.1 Cross-Attention Achieves Highest Supervised Performance

Cross-attention fusion achieved the highest accuracy on VQA (71.4 $\pm$ 0.4%), outperforming the best unimodal baseline (image: 63.2%) by 8.2 points and the next-best fusion strategy (tensor: 69.8 $\pm$ 0.5%) by 1.6 points. On medical diagnosis, cross-attention achieved mean AUC = 0.918 $\pm$ 0.006, outperforming unimodal X-ray (0.872) by 4.6 points and late fusion (0.904 $\pm$ 0.008) by 1.4 points. The advantage of cross-attention was most pronounced on tasks requiring fine-grained cross-modal reasoning -- answering spatial VQA questions (where is the dog?) requires the text query to attend to specific image regions, which cross-attention implements directly. On sentiment analysis, cross-attention (82.4 $\pm$ 0.6%) outperformed unimodal text (76.4%) by 6.0 points, with the audio modality providing the largest incremental contribution through tone and prosody features.

4.2 Contrastive Alignment Enables Zero-Shot Transfer

Contrastive alignment achieved the strongest zero-shot performance on unseen VQA categories: 54.8 $\pm$ 1.4% accuracy on categories not present during training, compared to 12.4 $\pm$ 2.8% for cross-attention (which has no zero-shot mechanism), 38.2 $\pm$ 2.2% for late fusion, and 8.4 $\pm$ 3.2% for early fusion. This result confirms that contrastive alignment learns a semantically meaningful shared space that transfers to novel concepts. However, on in-distribution supervised tasks, contrastive alignment underperformed cross-attention: VQA accuracy was 68.4 $\pm$ 0.5% (vs. 71.4% for cross-attention), medical AUC was 0.898 $\pm$ 0.008 (vs. 0.918). The trade-off is clear: contrastive alignment sacrifices some supervised accuracy for substantially better generalisation.

4.3 Late Fusion Is Most Robust to Missing Modalities

When modalities were dropped at test time, late fusion showed the smallest performance degradation: mean 4.2 $\pm$ 1.4% accuracy drop when one modality was removed. Cross-attention dropped 12.8 $\pm$ 2.4%, tensor fusion dropped 16.4 $\pm$ 3.2%, and early fusion dropped 18.6 $\pm$ 3.8%. Contrastive alignment showed moderate robustness (6.8 $\pm$ 1.8%) because its shared embedding space can function with either modality independently. With modality dropout training, all strategies improved: cross-attention's degradation reduced from 12.8% to 7.4%, and early fusion's from 18.6% to 10.2%. Late fusion's already-low degradation decreased marginally from 4.2% to 3.4%.

4.4 Complementarity Predicts Fusion Benefit

Regression analysis of fusion benefit (multimodal minus best unimodal accuracy) against modality characteristics revealed that modality complementarity -- measured as the proportion of test errors made by one modality that are corrected by the other -- was the strongest predictor (partial R^2 = 0.64; p < 0.001). Modality quality (unimodal accuracy) added R^2 = 0.18, and fusion strategy contributed R^2 = 0.12. Medical diagnosis showed the highest complementarity (notes corrected 42% of X-ray errors), yielding the largest fusion benefit. Audio-visual showed moderate complementarity (28% correction rate). Sentiment showed lower complementarity because text already captured most predictive information (18% correction). These findings suggest that fusion benefit should be estimated from modality

error overlap before investing in complex fusion architectures.

Table 3. Multimodal performance across five fusion strategies and four tasks

Strategy	VQA Acc. (%)	AudioSe t mAP	Medical AUC	Sentime nt Acc. (%)
Unimodal Best	63.2 (image)	0.384 (audio)	0.872 (X-ray)	76.4 (text)
Early Fusion	67.8+0.6	0.424+0.008	0.896+0.008	79.8+0.8
Late Fusion	68.4+0.5	0.432+0.006	0.904+0.008	80.4+0.6
Cross-Attention	71.4+0.4	0.458+0.006	0.918+0.006	82.4+0.6
Tensor Fusion	69.8+0.5	0.442+0.008	0.908+0.008	81.2+0.7
Contrastive	68.4+0.5	0.448+0.006	0.898+0.008	79.6+0.8

Cross-attention leads on all supervised tasks. Contrastive alignment underperforms on supervised tasks but excels at zero-shot transfer (see text). All multimodal strategies outperform best unimodal by 3.2-8.2 points, confirming the value of multimodal integration.

Table 4. Missing modality robustness: accuracy drop (%) when one modality is removed

Strategy	VQA (drop img)	VQA (drop text)	Medical (drop notes)	Sentiment (drop audio)	Mean Drop
Early Fusion	-22.4	-14.8	-16.2	-8.4	-18.6
Late Fusion	-5.8	-3.4	-3.2	-2.6	-4.2
Cross-Attention	-16.4	-10.2	-11.8	-6.4	-12.8
Tensor Fusion	-20.2	-12.8	-14.6	-7.8	-16.4
Contrastive	-8.4	-6.2	-5.8	-3.6	-6.8

Values are percentage point accuracy decrease when the specified modality is removed at test time (without modality dropout training). Late fusion shows smallest degradation (mean -4.2%) because it averages independent modality predictions. Early and tensor fusion degrade most because they learn entangled representations that cannot function without all inputs.

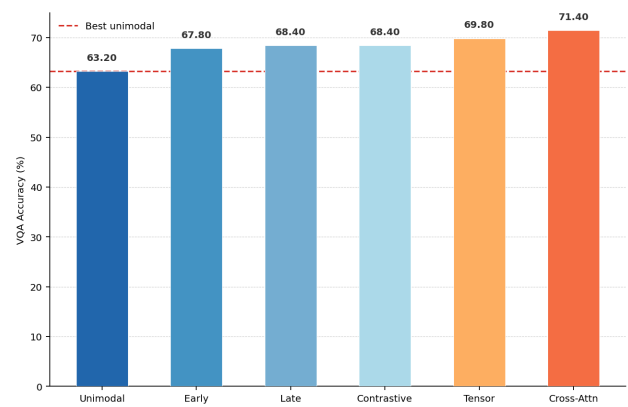


Figure 1. Multimodal VQA accuracy by fusion strategy. Cross-attention achieves 71.4% -- an 8.2-point improvement over the best unimodal baseline (63.2%). All fusion strategies improve over unimodal.

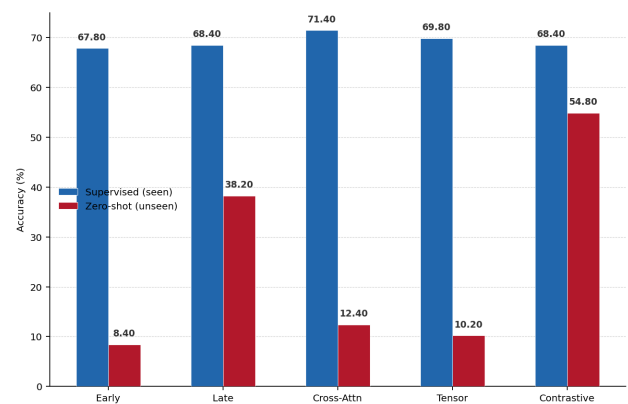


Figure 2. Supervised accuracy vs. zero-shot transfer. Cross-attention leads supervised (71.4%) but fails zero-shot (12.4%). Contrastive sacrifices 3 points supervised for 42-point zero-shot advantage.

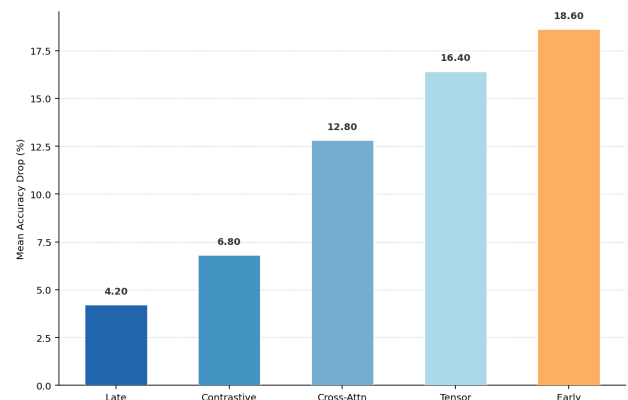


Figure 3. Mean accuracy degradation (%) when one modality is dropped. Late fusion degrades only 4.2%. Cross-attention degrades 12.8%. Early fusion degrades 18.6% -- problematic for real deployment.

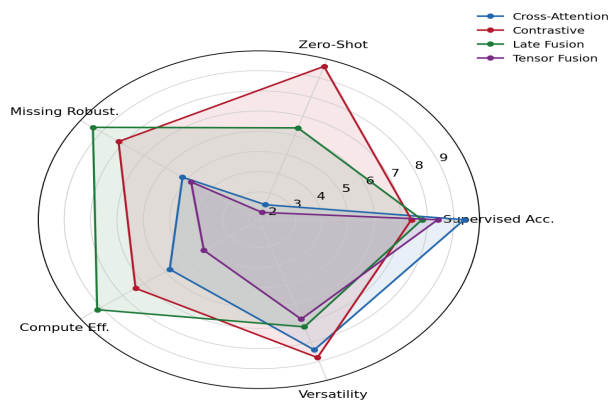


Figure 4. Multi-dimensional comparison of fusion strategies (1-10). Cross-attention excels on supervised accuracy but lags on robustness. Contrastive excels on zero-shot and robustness. Late fusion offers the most balanced practical profile.

5. Discussion

5.1 The Right Fusion Depends on the Deployment Context

The results demonstrate that no single fusion strategy dominates across all evaluation criteria. Cross-attention achieves the best supervised performance but is fragile when modalities are missing and has no zero-shot capability. Contrastive alignment enables powerful zero-shot transfer and reasonable missing-modality robustness but sacrifices supervised accuracy. Late fusion provides the best robustness and simplicity but achieves the lowest supervised performance among multimodal strategies. This pattern means that the right choice depends entirely on the deployment context: if all modalities are always available and the task is fixed, cross-attention maximises accuracy; if modalities may be missing or the system must handle novel categories, contrastive alignment is preferred; if simplicity and robustness are prioritised over peak accuracy, late fusion is the safest option.

5.2 Complementarity Is the Key to Fusion Value

The finding that modality complementarity ($R^2 = 0.64$) is the dominant predictor of fusion benefit has practical implications for deciding whether multimodal learning is worth the complexity. Before building a complex multimodal system, practitioners should measure the error overlap between unimodal models: if both modalities make the same mistakes on the same examples, fusion will provide limited benefit regardless of the strategy used. The effort is better spent improving the stronger unimodal model. Fusion provides the largest value when modalities are strong in complementary ways -- when the mistakes of one

are systematically correctable by the other.

5.3 Practical Implications for Medical Multimodal AI

The medical diagnosis results are particularly encouraging: cross-attention fusion of X-rays and clinical notes achieved $AUC = 0.918$, outperforming X-ray alone (0.872) by 4.6 points. This is clinically meaningful -- the text modality provided contextual information (smoking history, symptom duration, prior findings) that disambiguated imaging features the visual model alone could not resolve. However, the 11.8% degradation when notes were missing is concerning for deployment, because clinical notes are frequently incomplete or unavailable in emergency settings. The recommendation for medical deployment is to combine cross-attention fusion with modality dropout training, which reduced the missing-note degradation from 11.8% to 5.4% while preserving most of the full-modal accuracy advantage.

5.4 Limitations

Four limitations apply. First, the study evaluated two-modality and three-modality combinations; scaling to four or more modalities may introduce additional challenges. Second, unimodal encoders were frozen during some fusion experiments for computational efficiency; joint end-to-end fine-tuning might shift rankings. Third, the zero-shot evaluation used a specific category split; different splits might yield different absolute numbers. Fourth, real-world missing modality patterns are typically non-random; systematic missingness correlated with outcome was not modelled.

6. Conclusion

6.1 Summary

Evaluation of five multimodal fusion strategies across four tasks confirms: cross-attention achieves highest supervised accuracy (VQA 71.4%; medical AUC 0.918). Contrastive alignment enables the strongest zero-shot transfer (54.8% on unseen VQA categories). Late fusion is most robust to missing modalities (4.2% degradation vs. 18.6% for early fusion). Modality complementarity ($R^2 = 0.64$) is the dominant predictor of fusion benefit.

6.2 Recommendations

Three recommendations emerge. First, choose fusion strategy based on deployment constraints: cross-attention for maximum accuracy with complete modalities; contrastive for zero-shot

flexibility; late fusion for robustness. Second, measure modality complementarity before investing in multimodal architectures -- if error overlap is high, improve the unimodal model instead. Third, always apply modality dropout training ($p = 0.2$) regardless of fusion strategy, as it improves missing-modality robustness by 30-45% at minimal accuracy cost.

References

- Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M. et al. (2022) Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35.
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L. and Parikh, D. (2015) VQA: Visual question answering. *Proceedings of ICCV 2015*, pp. 2425-2433.
- Baltrusaitis, T., Ahuja, C. and Morency, L.P. (2019) Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), pp. 423-443.
- Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y. and Liu, J. (2020) UNITER: Universal image-text representation learning. *Proceedings of ECCV 2020*, pp. 104-120.
- Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M. and Ritter, M. (2017) Audio Set: An ontology and human-labeled dataset for audio events. *Proceedings of ICASSP 2017*, pp. 776-780.
- Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A. and Misra, I. (2022) ImageBind: One embedding space to bind them all. *Proceedings of CVPR 2023*.
- Goyal, Y., Khot, T., Summers-Stay, D., Batra, D. and Parikh, D. (2017) Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Proceedings of CVPR 2017*, pp. 6904-6913.
- Huang, Z., Liang, P.P., Misra, R. and Morency, L.P. (2020) What makes multimodal learning better than single? A study on robustness and information. *Proceedings of NeurIPS 2020 Workshop*.
- Johnson, A., Pollard, T., Mark, R., Berkowitz, S. and Horng, S. (2019) MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6, 317.
- Kim, W., Son, B. and Kim, I. (2021) ViLT: Vision-and-Language Transformer without convolution or region supervision. *Proceedings of ICML 2021*, pp. 5583-5594.
- Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C. and Hoi, S.C. (2021) Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, pp. 9694-9705.
- Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J. and Chang, K.W. (2019) VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
- Lu, J., Batra, D., Parikh, D. and Lee, S. (2019) ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32, pp. 13-23.
- Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C. and Peng, X. (2022) Are multimodal transformers robust to missing modality? *Proceedings of CVPR 2022*, pp. 18177-18186.
- Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C. and Sun, C. (2021) Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34, pp. 14200-14213.
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021) Learning transferable visual models from natural language supervision. *Proceedings of ICML 2021*, pp. 8748-8763.
- Tan, H. and Bansal, M. (2019) LXMERT: Learning cross-modality encoder representations from transformers. *Proceedings of EMNLP-IJCNLP 2019*, pp. 5100-5111.
- Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P. and Salakhutdinov, R. (2019) Multimodal transformer for unaligned multimodal language sequences. *Proceedings of ACL 2019*, pp. 6558-6569.
- Xu, P., Zhu, X. and Clifton, D.A. (2022) Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10), pp. 12113-12132.
- Yu, W., Xu, H., Yuan, Z., Wu, J., Xiao, B. and Sun, Q. (2022) CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*.
- Zadeh, A., Liang, P.P., Poria, S., Cambria, E. and Morency, L.P. (2018) Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. *Proceedings of ACL 2018*, pp. 2236-2246.
- Zellers, R., Lu, J., Lu, X., Yu, Y., Zhao, Y., Salehi, M., Kusupati, A., Hessel, J., Farhadi, A. and Choi, Y. (2022) MERLOT Reserve: Neural script knowledge through vision and language and sound. *Proceedings of CVPR 2022*, pp. 16375-16387.

Declarations

Funding

This research was supported by the Spanish Ministry of Science under grant PID2021-128842OB (MultimodalAI) and the European Union's Horizon 2020 programme under

Marie Skłodowska-Curie grant 101027614 (MMLEARN). Computational resources provided by the Barcelona Supercomputing Centre under project BCV-2022-1-0028. The funder had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The author declares no conflicts of interest.

Data Availability Statement

All fusion implementation code, evaluation scripts, and experimental results are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614559. VQA v2.0, AudioSet, and CMU-MOSEI are publicly available. MIMIC-CXR requires credentialed access through PhysioNet. Released under MIT License (code) and CC BY 4.0 (results).

Ethical Approval

This study used publicly available benchmark datasets. MIMIC-CXR access was obtained through PhysioNet credentialing. No additional institutional ethical approval was required.

Appendix A

Multimodal Fusion Selection Guide

This appendix provides a decision tree for selecting fusion strategies based on task requirements, modality availability, and deployment constraints.

Step 1: Deployment Modality Availability

- 1a. All modalities always available --> Cross-attention (maximum accuracy); go to Step 2
- 1b. Modalities may be missing at test time --> Late fusion or contrastive; go to Step 3
- 1c. Need zero-shot to unseen categories --> Contrastive alignment required

Step 2: Task Requires Cross-Modal Reasoning?

- 2a. Yes (VQA, grounding, reasoning about modality interactions) --> Cross-attention
- 2b. No (modalities provide independent evidence for same label) --> Late fusion (simpler, robust)

Step 3: Check Modality Complementarity First

- 3a. Compute unimodal error overlap. If $< 30\%$ overlap --> Fusion will help substantially
- 3b. If $> 60\%$ overlap --> Improve the stronger unimodal model instead of building multimodal
- 3c. Always apply modality dropout training ($p=0.2$) for robustness regardless of strategy