MANTECH Publications

Journal of Converging Technologies in Intelligent Computing
Volume 1, Issue 3, September-December 2025

Integration of Natural Language Processing and Computer Vision for Advanced Multimodal Intelligent Systems: A Systematic Review of Architectures, Benchmarks, and Applications

Arnab K. Dutta[1], Sreyashi P. Bhattacharya[2], Rajdeep S. Biswas[3], Mousumi D. Halder[4]

ABSTRACT

Natural Language Processing (NLP) and Computer Vision (CV) have long evolved as separate disciplines, yet most real-world intelligent systems must interpret images and language together. Recent advances in Transformer architectures, contrastive pre-training, and large language models have enabled a rapid convergence of the two fields into multimodal intelligent systems capable of visual question answering, captioning, retrieval, grounding, and instruction-following dialogue. This review-based study synthesises 30 peer-reviewed and widely cited publications spanning 2015–2024 and organises them through a four-dimensional analytical framework: fusion strategy, pre-training objective, data scale, and downstream performance. A taxonomy of five fusion paradigms is proposed, and benchmark results reported by the original authors are compared on a like-for-like basis. The synthesis shows that VQAv2 test-dev accuracy rose from 65.32% for region-based attention models to 82.0% for few-shot Flamingo-80B, a gain of 16.68 percentage points, while contrastive models reach roughly 76% zero-shot ImageNet accuracy without task-specific labels. Parameter-efficient bridging of frozen encoders and language models emerges as the dominant design trend, whereas object hallucination and weak college-level reasoning remain unresolved reliability barriers. Application mapping across healthcare, assistive technology, autonomous mobility, e-commerce, and education identifies deployment constraints, and a research agenda for low-resource, multilingual, and edge-deployable multimodal systems is outlined.

Page 174-188 © MANTECH PUBLICATIONS 2025. All Rights Reserved

KEYWORDS: *Multimodal learning, Natural language processing, Computer vision, Vision-language models, Transformers, Contrastive pre-training, Visual question answering, Large language models, Image captioning, Hallucination*

INTRODUCTION

Human understanding of the world is intrinsically multimodal: people read a road sign while looking at traffic, interpret a radiograph with the help of a clinical note, and describe a photograph in a sentence. For several decades, however, machine intelligence treated these abilities separately. Computer Vision relied on convolutional networks that map pixels to categorical labels [1], whereas Natural Language Processing built increasingly powerful sequence models over discrete tokens. The Transformer architecture [2] changed this situation by providing a single attention-based computational block for sequences. Bidirectional pre-training demonstrated that general-purpose language representations could be transferred to many tasks [3], and the Vision Transformer showed that an image split into patches could be processed by the same architecture with accuracy competitive with convolutional networks [4]. This architectural convergence is the technical foundation of modern multimodal intelligent systems.

Despite this progress, three problems restrict the dependable integration of language and vision. First, the two modalities have very different statistical structure, so aligning dense visual features with discrete, compositional text remains a representation problem rather than a simple concatenation problem [5]. Second, the best-performing systems depend on web-scale image–text corpora of hundreds of millions to billions of pairs [6, 7], which are costly to train on and poorly representative of many languages and regions. Third, fluent generation does not guarantee visual faithfulness; models may describe objects that are absent from an image [8, 9]. A clear problem statement therefore emerges: how should NLP and CV components be combined, trained, and evaluated so that multimodal systems are accurate, efficient, and trustworthy enough for practical deployment?

This review addresses that question through a structured comparative synthesis. Its contributions are a taxonomy of fusion strategies, a consolidated comparison of benchmark results reported for representative systems, a mapping of technical choices to application constraints, and an agenda for future research. The engineering significance lies in helping

Page 174-188 © MANTECH PUBLICATIONS 2025. All Rights Reserved

practitioners select an architecture that matches their data, compute, and reliability requirements instead of defaulting to the largest available model.

LITERATURE REVIEW

Early Image–Language Models and Benchmarks

Early multimodal systems coupled a convolutional image encoder with a recurrent language decoder. The encoder–decoder captioning model of Vinyals et al. generated sentences conditioned on a single global image vector [10], while the VQA benchmark introduced open-ended question answering about images and established a task requiring both recognition and reasoning [11]. Large annotated resources such as MS COCO Captions, with over 1.5 million captions for roughly 330,000 images, supplied the supervision necessary for training and evaluation [12]. Because early VQA models exploited language priors, a balanced second version of the dataset was released to force genuine use of visual evidence [13]. Attention over object-detector regions further improved both captioning and VQA by letting the model focus on salient entities [14].

Transformer-Based Vision-Language Pre-training

Following BERT, several studies pre-trained joint encoders on paired image–text data. ViLBERT processes visual and textual streams separately and exchanges information through co-attentional layers [15], whereas LXMERT uses separate object-relationship, language, and cross-modality encoders trained with five pre-training tasks [16]. Both depend on pre-computed region features, which makes inference slow. ViLT removed the detector and embedded image patches linearly, treating vision and text in a single shared Transformer and thereby reducing latency considerably, albeit with some loss of accuracy relative to region-based competitors [17].

Contrastive Language-Image Pre-training

Contrastive learning reframed the problem as aligning two encoders in a shared embedding space. CLIP was trained on 400 million image–text pairs collected from the web and could classify images zero-shot by comparing them with text prompts, matching the original ResNet-50 on ImageNet without using its training labels [6]. ALIGN showed that a simple dual encoder trained on 1.8 billion noisy alt-text pairs could achieve comparable or better transfer, indicating that scale can compensate for noise [7]. The same alignment principle was adapted to medicine,

Page 174-188 © MANTECH PUBLICATIONS 2025. All Rights Reserved

where paired chest radiographs and reports provided a source of supervision that avoids costly manual labels [18].

Generative and LLM-Based Multimodal Systems

Generative formulations followed. BLIP combined understanding and generation in one model and introduced a captioner–filter procedure that bootstraps cleaner synthetic captions from noisy web text [19]. Flamingo interleaved frozen language-model layers with gated cross-attention to a visual resampler, enabling few-shot adaptation from a handful of examples [20]. BLIP-2 reduced training cost further by bridging a frozen image encoder and a frozen language model through a lightweight Querying Transformer [21], and LLaVA applied visual instruction tuning, projecting visual features into a language model using machine-generated instruction data [22]. These systems build upon the few-shot capability of large language models [23] and the wider paradigm of foundation models [24]. In the opposite direction, text-conditioned image synthesis with latent diffusion [25] and CLIP-latent priors [26] shows that language can also control visual generation.

Reliability and Evaluation

As capability has grown, evaluation has become central. A broad survey classifies hallucination in generated text as intrinsic or extrinsic and discusses mitigation [8], and polling-based evaluation shows that vision-language models frequently hallucinate objects, especially those that commonly co-occur with others in training data [9]. The MMMU benchmark, built from college-level questions across six disciplines, reported that even the strongest evaluated model, GPT-4V, reached only 55.7% accuracy [27]. A foundational survey organises the field into representation, translation, alignment, fusion, and co-learning challenges [5], which informs the framework adopted below.

RESEARCH GAP

The literature reveals four gaps. First, results are reported under heterogeneous pre-training data, backbone sizes, and evaluation settings, so claims of superiority are rarely comparable across model families. Second, few studies connect architectural choices, such as late fusion versus cross-attention, with practical constraints such as latency, trainable-parameter budget, and data availability. Third, reliability issues, including hallucination and weak expert-level reasoning, are typically treated separately from accuracy rather than as a joint design criterion.

Page 174-188 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Fourth, evidence on low-resource languages and regional contexts, including Indic languages, is considerably thinner than evidence for English. This review addresses these gaps through a unified, review-based comparison.

OBJECTIVES

- To develop a taxonomy of fusion strategies used to integrate NLP and CV components in multimodal intelligent systems.
- To compare reported benchmark performance of representative vision-language models under consistent metrics and to identify performance trends.
- To analyse the relationship between data scale, training objective, and trainable-parameter efficiency.
- To map technical capabilities and reliability limitations onto practical application domains.
- To formulate a prioritised research agenda for dependable, efficient, and multilingual multimodal systems.

METHODOLOGY

Review Design and Selection Criteria

This study is a review-based comparative synthesis, designed according to the structured guidelines for software-engineering and computing reviews proposed by Kitchenham and Charters [28] and reported in the spirit of the PRISMA 2020 statement [29]. A purposive corpus of 30 publications was assembled from major venues in machine learning, vision, and language processing, including CVPR, ICCV, ICML, NeurIPS, ICLR, NAACL, EMNLP, and IEEE TPAMI, covering 2015–2024. Publications were included if they (a) jointly modelled visual and textual data, (b) reported results on publicly available benchmarks, and (c) described the architecture in sufficient detail for classification. Unimodal studies and proprietary systems lacking a technical report were excluded. Because the corpus is purposive rather than exhaustive, it is intended to represent each architectural paradigm instead of enumerating every published model.

Analytical Framework

Every selected model was coded along four dimensions: (i) fusion strategy, describing where and how the modalities interact; (ii) pre-training objective, such as image–text contrastive learning, image–text matching, masked language modelling, or autoregressive language

Page 174-188 © MANTECH PUBLICATIONS 2025. All Rights Reserved

modelling; (iii) data scale, measured as the number of image–text pairs or images used in pre-training; and (iv) downstream performance on standard benchmarks. Fusion strategies were grouped into five paradigms, summarised in Table 1, extending the alignment and fusion categories of the multimodal learning taxonomy [5]. Figure 1 shows the generic reference architecture used to position each model within the pipeline.

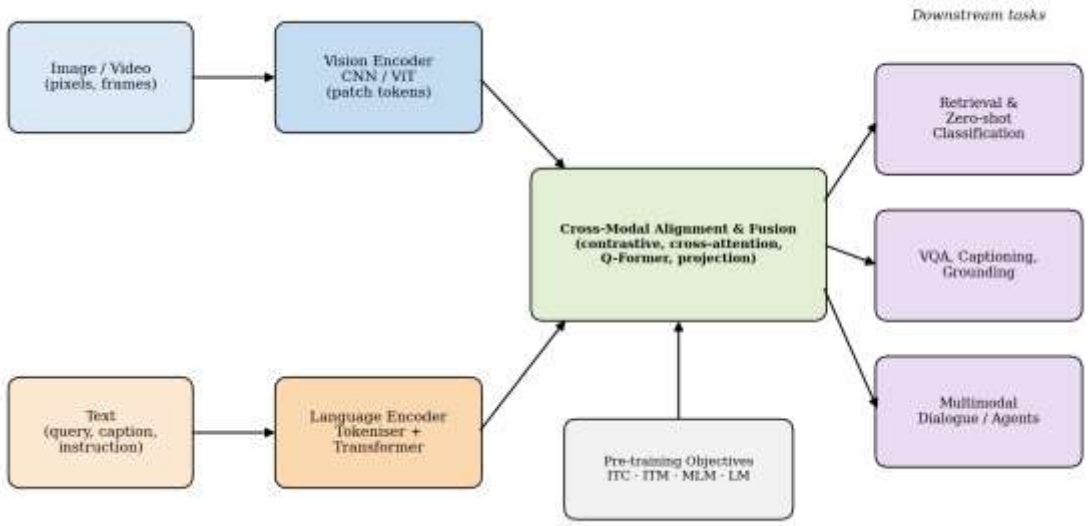


Figure 1. Generic reference architecture for integrating NLP and CV: modality-specific encoders, a cross-modal alignment and fusion module, shared pre-training objectives (ITC: image–text contrastive; ITM: image–text matching; MLM: masked language modelling; LM: language modelling), and downstream tasks

Table 1. Taxonomy of Fusion Strategies for Multimodal NLP–CV Systems

Fusion paradigm	Representative models	Strengths	Limitations
Dual-encoder contrastive	CLIP, ALIGN [6, 7]	Efficient retrieval; zero-shot transfer	Weak fine-grained reasoning
Two-stream co-attention	ViLBERT, LXMERT [15, 16]	Rich region–word interaction	Slow region features; high cost
Single-stream patch-level	ViLT [17]	Detector-free; fast inference	Lower accuracy than region models
Unified encoder–decoder	BLIP [19]	Understanding and generation in one model	Needs large, cleaned data
Frozen-LLM bridging	Flamingo, BLIP-2, LLaVA [20, 21, 22]	Few-shot ability; few trainable parameters	Hallucination; high inference cost

Page 174-188 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Comparison Protocol

Performance values were extracted exactly as reported by the original authors and were not re-executed. To limit unfair comparison, models were compared only on the same benchmark and metric: accuracy on the VQAv2 test-dev split for question answering [13] and zero-shot top-1 accuracy on ImageNet for contrastive models. Zero-shot and fine-tuned results were never merged into the same ranking, and differences are expressed in percentage points. Trends in data scale and trainable parameters were interpreted qualitatively, with explicit acknowledgement that backbones, data, and compute differ between studies.

RESULTS AND FINDINGS

The synthesis produced five principal findings. Figure 2 illustrates the historical progression, while Figure 3 and Table 2 consolidate the quantitative evidence.

Finding 1: the field progressed through four design phases, summarised in Figure 2. Region-feature and recurrent models gave way to fusion Transformers in 2019, contrastive and unified models in 2021–2022, and language-model-centred systems from 2022 onward.

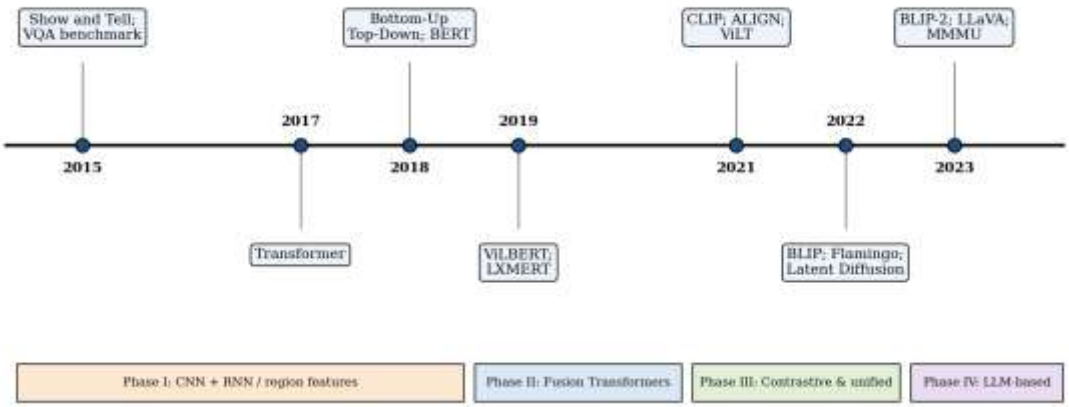


Figure 2. Timeline of milestones in the integration of NLP and CV, grouped into four design phases (2015–2023)

Finding 2: reported VQAv2 test-dev accuracy increased steadily with each phase. The region-based Bottom-Up Top-Down model achieved 65.32% [14], ViLBERT 70.55% [15], LXMERT 72.42% [16], and ViLT 71.26% [17]. BLIP raised the figure to 78.25% [19], and fine-tuned Flamingo-80B reached 82.0% [20]. The cumulative improvement from 65.32% to 82.0% is 16.68 percentage points, of which 12.93 points came from the move to unified generative pre-

Page 174-188 © MANTECH PUBLICATIONS 2025. All Rights Reserved

training between 2018 and 2022. ViLT trails LXMERT by 1.16 points but avoids the region detector, illustrating an accuracy–efficiency trade-off.

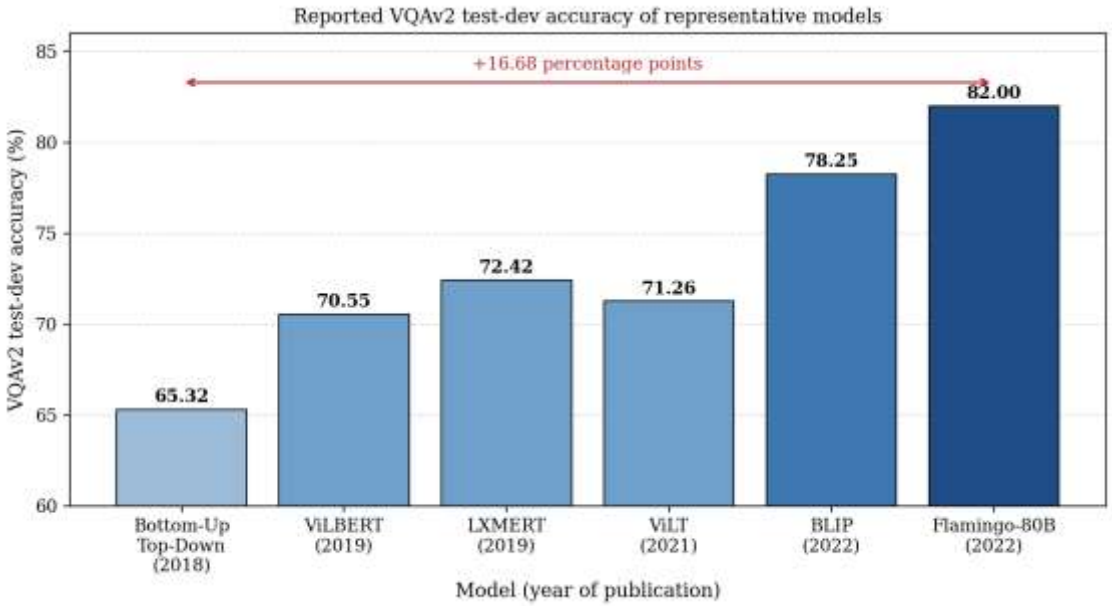


Figure 3. Reported VQAv2 test-dev accuracy (%) of representative vision-language models; values are taken from the original publications [14, 15, 16, 17, 19, 20]

Table 2. Comparison of Representative Models Using Results Reported in the Original Publications

Model	Year	Fusion paradigm	Pre-training data	Reported result
CLIP [6]	2021	Dual encoder	400 M image–text pairs	76.2% zero-shot ImageNet
ALIGN [7]	2021	Dual encoder	1.8 B noisy pairs	76.4% zero-shot ImageNet
ViLBERT [15]	2019	Two-stream	3.3 M (Conceptual Captions)	70.55% VQAv2 test-dev
LXMERT [16]	2019	Two-stream	9.18 M pairs	72.42% VQAv2 test-dev
ViLT [17]	2021	Single-stream	4 M images	71.26% VQAv2 test-dev
BLIP [19]	2022	Encoder–decoder	129 M images	78.25% VQAv2 test-dev

Page 174-188 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Model	Year	Fusion paradigm	Pre-training data	Reported result
Flamingo-80B [20]	2022	Frozen-LLM bridging	Web-scale interleaved data	82.0% VQAv2 test-dev (fine-tuned)

Finding 3: data scale shows diminishing returns within the contrastive family. ALIGN used 1.8 billion pairs, about 4.5 times the 400 million used by CLIP, yet its zero-shot ImageNet accuracy of 76.4% exceeds CLIP's 76.2% by only 0.2 points [6, 7]. Although the backbones differ, the result suggests that additional noisy data brings smaller gains than improved data quality; this is consistent with BLIP, whose caption-filtering strategy improved performance by cleaning web text rather than simply enlarging it [19].

Finding 4: parameter efficiency has become a primary design objective. BLIP-2 trains only a lightweight bridging module while freezing the image encoder and the language model, and its authors report outperforming Flamingo-80B by 8.7% on zero-shot VQAv2 with 54 times fewer trainable parameters [21]. LLaVA likewise adapts a pre-trained language model with a projection and instruction tuning rather than training from scratch [22]. Fine-tuning approaches based on low-rank adaptation extend the same principle to downstream specialisation [30].

Finding 5: reliability has not kept pace with accuracy. Despite strong benchmark scores, GPT-4V achieved 55.7% on the college-level MMMU benchmark [27], and polling-based evaluation shows systematic object hallucination in large vision-language models [9]. Accuracy on perceptual benchmarks therefore overstates readiness for expert or safety-critical use.

DISCUSSION

The findings indicate that integration succeeded where the interface between modalities became simpler and the pre-trained unimodal components became stronger. Contrastive alignment is attractive because it needs only a similarity objective and yields embeddings that support retrieval and zero-shot classification [6]. Cross-attention fusion remains preferable when fine-grained reasoning over regions or tokens is required [16, 15], but its cost motivated single-stream and bridged designs [17, 21]. The trend toward frozen backbones connected by small trainable modules suggests that future systems will be assembled from reusable unimodal foundations rather than trained end-to-end, which lowers the energy and hardware barrier for

Page 174-188 © MANTECH PUBLICATIONS 2025. All Rights Reserved

smaller institutions. For academic laboratories and regional engineering colleges in particular, training only a bridging module on a modest dataset is realistic, whereas pre-training a dual encoder on hundreds of millions of pairs is not. Open model checkpoints therefore function as shared infrastructure, and research effort is better directed at adaptation, evaluation, and data curation than at scale alone.

From an engineering viewpoint, the choice of architecture should follow deployment constraints. Retrieval and cataloguing favour dual encoders because image embeddings can be pre-computed and indexed; interactive assistants favour LLM-bridged models; latency-bound edge applications favour detector-free, single-stream encoders. The scientific significance of the review is the identification of a decoupling between benchmark accuracy and trustworthiness: the same models that approach 80% on VQAv2 remain vulnerable to hallucination and fail on expert-level reasoning [9, 27].

Practical applications are summarised in Table 3. In healthcare, image–report alignment reduces dependence on manual annotation [18]; in assistive technology, visual question answering describes surroundings for people with visual impairment [11]; in e-commerce and digital libraries, zero-shot retrieval enables search by natural language without category-specific training [6]; and in creative and educational tools, text-conditioned generation supports illustration and content design [25]. In each case, the consequence of an error determines the level of human oversight required.

Table 3. Application Domains, Technical Contributions, and Principal Challenges

Domain	NLP contribution	CV contribution	Key challenge
Healthcare [18]	Report and clinical-text encoding	Radiograph feature extraction	Privacy; error cost
Assistive technology [11]	Question understanding; answer generation	Scene and object recognition	Reliability; latency
E-commerce and libraries [6]	Natural-language query encoding	Product and image embeddings	Domain shift; bias
Education and design [25]	Prompt and instruction parsing	Image synthesis and layout	Faithfulness; copyright

Page 174-188 © MANTECH PUBLICATIONS 2025. All Rights Reserved

Domain	NLP contribution	CV contribution	Key challenge
Autonomous mobility [14]	Instruction and sign interpretation	Detection and attention over regions	Safety certification

LIMITATIONS

- The study is review-based: benchmark values are taken from original publications and were not re-run under a common training and evaluation protocol, so cross-paper differences may partly reflect backbone size, data, and compute.
- The corpus was selected purposively and is not exhaustive; the rapid release of new models, including proprietary systems without technical reports, means some recent advances are not represented.
- Public benchmarks are predominantly English-centred and may suffer from training-data overlap, which can inflate reported scores and limits conclusions about multilingual or regional performance.

FUTURE SCOPE

- Development of open, culturally representative image–text corpora in Indic and other low-resource languages, with standardised multilingual multimodal benchmarks.
- Hallucination-aware training objectives and joint accuracy–faithfulness evaluation protocols, combined with calibrated uncertainty estimation.
- Parameter-efficient adaptation and model compression for deployment on edge devices and in institutions with limited computing resources [30].
- Extension from static images to video, audio, and sensor streams, together with privacy-preserving and explainable multimodal learning for clinical use.

CONCLUSION

This review examined how NLP and CV have been integrated into multimodal intelligent systems and organised the literature into five fusion paradigms and four historical phases. Reported VQAv2 test-dev accuracy improved by 16.68 percentage points between region-based attention models and Flamingo-80B, while contrastive models achieved around 76% zero-shot ImageNet accuracy without task-specific labels. The decisive trends are the move toward simpler modality interfaces, the use of frozen pre-trained backbones, and an increasing emphasis on data quality over raw data volume. However, hallucination and limited expert-

Page 174-188 © MANTECH PUBLICATIONS 2025. All Rights Reserved

level reasoning show that benchmark accuracy alone is an insufficient indicator of readiness. Future work should prioritise multilingual data, faithfulness-aware evaluation, and efficient adaptation so that multimodal systems become both capable and dependable in real-world practice.

REFERENCES

1. He K, Zhang X, Ren S, Sun J. *Deep residual learning for image recognition.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; 770–778.
2. Vaswani A, Shazeer N, Parmar N, et al. *Attention is all you need.* Advances in Neural Information Processing Systems. 2017; 30: 5998–6008.
3. Devlin J, Chang MW, Lee K, Toutanova K. *BERT: Pre-training of deep bidirectional transformers for language understanding.* Proceedings of NAACL-HLT. 2019; 4171–4186.
4. Dosovitskiy A, Beyer L, Kolesnikov A, et al. *An image is worth 16x16 words: Transformers for image recognition at scale.* International Conference on Learning Representations (ICLR). 2021.
5. Baltrušaitis T, Ahuja C, Morency LP. *Multimodal machine learning: A survey and taxonomy.* IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019; 41(2): 423–443.
6. Radford A, Kim JW, Hallacy C, et al. *Learning transferable visual models from natural language supervision.* Proceedings of the 38th International Conference on Machine Learning (PMLR). 2021; 139: 8748–8763.
7. Jia C, Yang Y, Xia Y, et al. *Scaling up visual and vision-language representation learning with noisy text supervision.* Proceedings of the 38th International Conference on Machine Learning (PMLR). 2021; 139: 4904–4916.
8. Ji Z, Lee N, Frieske R, et al. *Survey of hallucination in natural language generation.* ACM Computing Surveys. 2023; 55(12): 1–38.
9. Li Y, Du Y, Zhou K, Wang J, Zhao WX, Wen JR. *Evaluating object hallucination in large vision-language models.* Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2023; 292–305.

Page 174-188 © MANTECH PUBLICATIONS 2025. All Rights Reserved

10. Vinyals O, Toshev A, Bengio S, Erhan D. *Show and tell: A neural image caption generator.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015; 3156–3164.
11. Antol S, Agrawal A, Lu J, et al. *VQA: Visual question answering.* Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2015; 2425–2433.
12. Chen X, Fang H, Lin TY, et al. *Microsoft COCO Captions: Data collection and evaluation server.* arXiv preprint arXiv:1504.00325. 2015.
13. Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D. *Making the V in VQA matter: Elevating the role of image understanding in visual question answering.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017; 6904–6913.
14. Anderson P, He X, Buehler C, et al. *Bottom-up and top-down attention for image captioning and visual question answering.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018; 6077–6086.
15. Lu J, Batra D, Parikh D, Lee S. *ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.* Advances in Neural Information Processing Systems. 2019; 32: 13–23.
16. Tan H, Bansal M. *LXMERT: Learning cross-modality encoder representations from transformers.* Proceedings of EMNLP-IJCNLP. 2019; 5100–5111.
17. Kim W, Son B, Kim I. *ViLT: Vision-and-language transformer without convolution or region supervision.* Proceedings of the 38th International Conference on Machine Learning (PMLR). 2021; 139: 5583–5594.
18. Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. *Contrastive learning of medical visual representations from paired images and text.* Proceedings of Machine Learning for Healthcare (PMLR). 2022; 182: 2–25.
19. Li J, Li D, Xiong C, Hoi S. *BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation.* Proceedings of the 39th International Conference on Machine Learning (PMLR). 2022; 162: 12888–12900.
20. Alayrac JB, Donahue J, Luc P, et al. *Flamingo: A visual language model for few-shot learning.* Advances in Neural Information Processing Systems. 2022; 35: 23716–23736.

Page 174-188 © MANTECH PUBLICATIONS 2025. All Rights Reserved

21. Li J, Li D, Savarese S, Hoi S. *BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.* Proceedings of the 40th International Conference on Machine Learning (PMLR). 2023; 202: 19730–19742.
22. Liu H, Li C, Wu Q, Lee YJ. *Visual instruction tuning.* Advances in Neural Information Processing Systems. 2023; 36.
23. Brown T, Mann B, Ryder N, et al. *Language models are few-shot learners.* Advances in Neural Information Processing Systems. 2020; 33: 1877–1901.
24. Bommasani R, Hudson DA, Adeli E, et al. *On the opportunities and risks of foundation models.* arXiv preprint arXiv:2108.07258. 2021.
25. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. *High-resolution image synthesis with latent diffusion models.* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022; 10684–10695.
26. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. *Hierarchical text-conditional image generation with CLIP latents.* arXiv preprint arXiv:2204.06125. 2022.
27. Yue X, Ni Y, Zhang K, et al. *MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI.* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024; 9556–9567.
28. Kitchenham B, Charters S. *Guidelines for performing Systematic Literature Reviews in Software Engineering (EBSE Technical Report).* Keele University and University of Durham. 2007.
29. Page MJ, McKenzie JE, Bossuyt PM, et al. *The PRISMA 2020 statement: An updated guideline for reporting systematic reviews.* BMJ. 2021; 372: n71.
30. Hu EJ, Shen Y, Wallis P, et al. *LoRA: Low-rank adaptation of large language models.* International Conference on Learning Representations (ICLR). 2022.

Page 174-188 © MANTECH PUBLICATIONS 2025. All Rights Reserved

***Author for Correspondence:**
Arnab K. Dutta
E-mail: arnab.dutta.cse@gmail.com

[1]Associate Professor, Department of Computer Science and Engineering, Dr. Sudhir Chandra Sur Institute of Technology and Sports Complex, Kolkata, West Bengal
[2]Assistant Professor, Department of Computer Science and Engineering, Dr. Sudhir Chandra Sur Institute of Technology and Sports Complex, Kolkata, West Bengal
[3]Assistant Professor, Department of Computer Science and Engineering, Dr. Sudhir Chandra Sur Institute of Technology and Sports Complex, Kolkata, West Bengal
[4]Research Scholar, Department of Computer Science and Engineering, Dr. Sudhir Chandra Sur Institute of Technology and Sports Complex, Kolkata, West Bengal

Received Date: October 07, 2025
Accepted Date: October 19, 2025
Published Date: October 30, 2025

Citation: Arnab K. Dutta, Sreyashi P. Bhattacharya, Rajdeep S. Biswas, Mousumi D. Halder. Integration of Natural Language Processing and Computer Vision for Advanced Multimodal Intelligent Systems: A Systematic Review of Architectures, Benchmarks, and Applications. Journal of Converging Technologies in Intelligent Computing. 2025; 1(3): 174-188p.

Page 174-188 © MANTECH PUBLICATIONS 2025. All Rights Reserved