Transactions on Computational and Scientific Methods | Vo. 5, No. 10, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

# From Vision to Reasoning: Leveraging Deep Learning for Enhancing Large Language Models in Multimodal Understanding

**Linnea Forsythe**
Eastern Washington University, Cheney, USA
lf43@ewu.edu

**Abstract:** In recent years, the integration of deep learning and large language models (LLMs) has become a transformative force in artificial intelligence, driving advances in multimodal understanding, reasoning, and human-computer interaction. While LLMs exhibit strong linguistic and reasoning capabilities, their perception of non-textual modalities such as images, videos, and signals remains limited. This paper proposes a unified framework named DeepVision-Reasoner, which leverages deep neural architectures to enhance the multimodal reasoning capacity of LLMs. The framework integrates a vision encoder based on convolutional and transformer-based representations with a large language decoder, enabling the model to learn from both visual and textual sources in an end-to-end manner. The proposed method introduces a dual-stage alignment process that harmonizes visual embeddings with linguistic tokens through a shared latent space and an adaptive cross-attention mechanism. Extensive experiments across visual question answering, caption generation, and image-grounded reasoning demonstrate that the proposed model significantly outperforms baseline multimodal LLMs in accuracy, coherence, and semantic grounding. Moreover, the model exhibits robust generalization under zero-shot settings, highlighting the synergy between deep learning feature extraction and large-scale generative reasoning. This study contributes to the ongoing convergence between perceptual deep networks and cognitive-level language models, paving the way for more unified, human-like artificial intelligence systems.

**Keywords:** Deep Learning, Large Language Models, Multimodal Reasoning, Vision-Language Alignment, Artificial Intelligence

## 1. Introduction

In the evolving landscape of artificial intelligence, deep learning and large language models (LLMs) have independently demonstrated remarkable breakthroughs in perception and cognition, respectively. Deep learning architectures, such as convolutional neural networks (CNNs), vision transformers (ViTs), and graph neural networks (GNNs), have enabled machines to process and understand complex visual, spatial, and temporal data at levels once unattainable. Conversely, LLMs-represented by models like GPT, PaLM, and LLaMA-have redefined natural language understanding and reasoning, enabling context-aware generation, semantic alignment, and logical inference across diverse linguistic domains. However, despite these advancements, a fundamental divide persists between visual perception and linguistic reasoning. Deep learning models excel in extracting high-dimensional, fine-grained features, yet lack the interpretative and contextual depth that LLMs possess. On the other hand, LLMs exhibit strong symbolic reasoning and compositional understanding but remain limited in grounding abstract linguistic representations in visual or sensory experiences. Bridging this cognitive gap represents one of the most pressing frontiers in modern AI research.

The rise of multimodal AI-integrating text, vision, and other modalities-has underscored the need for unified frameworks that harmonize deep feature learning with language-based reasoning. Existing approaches, such as CLIP, Flamingo, and GPT-4V, attempt to connect image encoders with language models through contrastive pretraining or attention-based fusion. While these frameworks achieve commendable results in image-text alignment, they often rely on loosely coupled architectures where the visual and linguistic subsystems remain distinct. This architectural decoupling limits deep semantic fusion, resulting in shallow cross-modal reasoning and brittle performance in tasks requiring compositional inference or zero-shot adaptation. Moreover, most current models are highly data-dependent and fail to generalize efficiently when confronted with unseen multimodal configurations, thereby restricting their scalability across domains such as robotics, medical imaging, and scientific analysis.

To address these challenges, this paper introduces DeepVision-Reasoner, an end-to-end framework designed to seamlessly integrate deep visual representations into large language models for enhanced multimodal understanding and reasoning. Unlike conventional pipelines that treat vision encoders as static feature extractors, our approach embeds the deep learning component dynamically within the LLM's architecture, enabling continuous information exchange between perceptual and linguistic modules. The model utilizes a dual-stream attention mechanism where visual embeddings and linguistic tokens are projected into a shared latent space through adaptive cross-modal transformers. This enables the system to generate contextually grounded textual outputs that accurately reflect the semantics of the visual input while maintaining linguistic fluency and logical consistency. By jointly optimizing vision-language objectives, DeepVision-Reasoner learns to align heterogeneous modalities at both the representational and reasoning levels.

The proposed framework represents a step toward what we define as cognitive fusion: the unification of perception and reasoning within a single, coherent neural architecture. Such fusion not only improves interpretability and generalization but also offers a pathway toward AI systems capable of holistic understanding-systems that can "see" and "think" simultaneously. This paradigm holds vast potential in real-world applications, from autonomous driving and medical diagnostics to education and creative design. Furthermore, integrating deep learning and LLMs in this manner allows the model to self-adapt through few-shot and zero-shot learning, reducing dependence on labeled multimodal datasets. Ultimately, the goal of this study is to advance multimodal intelligence from mere perception-based recognition to genuine reasoning, bridging the gap between low-level sensory data and high-level cognitive processes.

## 2. Related Work

The integration of deep learning and large language models (LLMs) in multimodal reasoning represents an interdisciplinary convergence that spans computer vision, natural language processing, and representation learning. Early work in visual understanding, led by convolutional neural networks (CNNs), laid the foundation for perceptual intelligence through hierarchical feature extraction. Models such as AlexNet [1], VGG [2], and ResNet [3] established deep feature hierarchies that enabled machines to recognize visual patterns with remarkable precision. Later, the emergence of transformer-based architectures, particularly the Vision Transformer (ViT) [4], demonstrated that self-attention mechanisms could outperform convolutional models in image classification and object detection tasks by capturing long-range dependencies. These developments in deep visual representation learning became critical enablers for cross-modal alignment with LLMs.

In parallel, large language models underwent a rapid evolution driven by advances in self-supervised learning and large-scale pretraining. The Transformer architecture [5] revolutionized sequence modeling, leading to the creation of GPT [6], BERT [7], and T5 [8], which established the foundation for modern generative reasoning systems. These models demonstrated unprecedented abilities in context understanding, abstraction, and logical inference. However, despite their linguistic fluency, traditional LLMs were constrained by their unimodal nature-they processed symbolic information without grounding in perceptual data. This limitation inspired research into multimodal models capable of fusing textual and non-textual modalities.

Recent breakthroughs in multimodal learning have sought to unify visual and linguistic representations through contrastive learning, cross-attention, or joint pretraining. The CLIP model [9], developed by OpenAI, was a pioneering effort that learned a shared embedding space for images and text using contrastive language–image pretraining. This approach allowed zero-shot classification by aligning visual and textual representations. Following CLIP, models such as ALIGN [10] and BLIP [11] further enhanced visual-language pretraining through large-scale web data and improved fusion strategies. However, while these models achieved high performance in recognition and retrieval tasks, they often fell short in compositional reasoning, logical inference, and contextual understanding due to their reliance on shallow alignment mechanisms.

More recently, hybrid architectures have emerged that integrate vision encoders directly into LLMs, giving rise to multimodal large language models (MLLMs). Examples include Flamingo [12], which introduced a gated cross-attention mechanism for vision–language fusion, and GPT-4V [13], which extended generative language modeling to visual inputs. Similarly, PaLM-E [14] connected pretrained LLMs with vision transformers to create embodied reasoning systems capable of processing video and sensory input. Despite these advances, current multimodal LLMs remain limited by the weak coupling between perceptual and cognitive components, often depending on frozen image encoders that restrict bidirectional information flow. This architectural rigidity hinders joint optimization and results in incomplete semantic grounding between modalities.

From the perspective of representational learning, several works have explored aligning visual features with linguistic tokens in a shared latent space. For example, Li et al. [15] proposed a unified embedding model that projects both image and text features into a common manifold using a mutual information constraint, while Xu et al. [16] employed a semantic correlation module to enforce cross-modal consistency. However, most of these approaches assume static mappings between modalities and fail to adapt dynamically to contextual changes in multimodal data. In contrast, our proposed framework DeepVision-Reasoner introduces an adaptive alignment mechanism that jointly optimizes both modalities during training, ensuring continuous co-adaptation between perception and reasoning layers.

Another related line of research lies in cognitive-inspired multimodal reasoning. Recent efforts, such as VisualGPT [17] and LLaVA [18], have demonstrated that integrating visual encoders into generative language models enables coherent visual grounding during text generation. Nonetheless, these systems primarily rely on instruction tuning or caption-based datasets and do not fully exploit the potential of deep learning for perceptual enhancement. Our approach differs by embedding the deep learning model as a dynamic perceptual module within the LLM structure, allowing the system to refine visual understanding in tandem with language-based reasoning. This integration embodies a deeper form of neural synergy that mimics human cognition-perception feeding into reasoning and reasoning shaping perception.

In summary, while existing multimodal frameworks have made significant strides in bridging the visual–linguistic gap, challenges persist in achieving deep semantic fusion, dynamic co-adaptation, and reasoning-grounded understanding. The proposed DeepVision-Reasoner aims to overcome these limitations through an end-to-end trainable architecture that unifies perceptual learning with cognitive reasoning, enabling more holistic and generalizable multimodal intelligence.

## 3. Proposed Approach

The proposed DeepVision-Reasoner framework is designed to seamlessly integrate deep visual representation learning with large language model reasoning within a unified architecture. The central objective is to enable mutual enhancement between perception and cognition - where deep learning modules extract semantically rich features from visual inputs, and the LLM refines reasoning through contextual language modeling. The system follows an encoder–decoder paradigm, in which the visual encoder transforms raw image data into dense feature embeddings, and the LLM-based decoder performs cross-

modal reasoning and textual generation grounded in those embeddings. Figure 1 illustrates the overall architecture of the proposed system, highlighting the visual encoder, cross-modal alignment layer, and language decoder that together form the end-to-end learning pipeline.

The architecture begins with a deep visual encoder $E_v(\cdot)$, based on a hybrid convolution-transformer backbone. Given an image $I$, the encoder extracts hierarchical representations $V = E_v(I) \in \mathbb{R}^{n \times d_v}$, where $n$ denotes the number of visual tokens and $d_v$ represents the embedding dimension. These tokens capture both local spatial structure and global semantic context through self-attention layers. To bridge the modality gap between visual and linguistic representations, an adaptive projection module $f_p(\cdot)$transforms $V$into a shared latent space compatible with the language model embedding dimension $d_l$:

$$Z_v = f_p(V) = \text{LayerNorm}(W_p V + b_p)$$

where $W_p \in \mathbb{R}^{d_l \times d_v}$and $b_p \in \mathbb{R}^{d_l}$are learnable parameters. This transformation ensures that visual and linguistic tokens coexist within a unified vector space, enabling cross-modal attention within the LLM layers.

The language reasoning module is implemented using a transformer-based LLM $E_l(\cdot)$, pretrained on large-scale text corpora. The textual input $T = \{t_1, t_2, \ldots, t_m\}$is first embedded into token representations $L = E_t(T) \in \mathbb{R}^{m \times d_l}$. The multimodal fusion then occurs via a cross-attention mechanism, where visual embeddings $Z_v$serve as key–value pairs and language embeddings $L$as queries. This attention operation allows the language model to condition its next-token prediction on both linguistic context and visual semantics:

$$\text{Attention}(L, Z_v) = \text{softmax}\left(\frac{L W_Q (Z_v W_K)^\top}{\sqrt{d_l}}\right)(Z_v W_V)$$

where $W_Q, W_K, W_V$are the projection matrices for query, key, and value, respectively. Through this cross-attention layer, *DeepVision-Reasoner* enables bidirectional interaction-language tokens attend to visual information, while visual features are refined via linguistic feedback during training.

The training objective combines a multimodal reasoning loss with a language generation loss. The overall optimization function is defined as:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{align} + \lambda_2 \mathcal{L}_{gen}$$

Here, $\mathcal{L}_{align}$enforces representational consistency between visual and textual embeddings via a contrastive alignment objective:

$$\mathcal{L}_{align} = -\log \frac{\exp(\text{sim}(Z_v, L^+)/\tau)}{\sum_{L^-} \exp(\text{sim}(Z_v, L^-)/\tau)}$$

where sim$(\cdot)$denotes cosine similarity, $\tau$is a temperature parameter, $L^+$represents the ground-truth paired caption, and $L^-$indicates negative samples. Meanwhile, $\mathcal{L}_{gen}$corresponds to the autoregressive language

modeling loss that maximizes the log-likelihood of generating the correct textual output conditioned on visual context:

$$\mathcal{L}_{gen} = -\sum_{t=1}^{m} \log P(t_t | t_{<t}, Z_v)$$

Jointly optimizing these objectives encourages the system to not only align multimodal representations but also reason coherently in the presence of visual grounding.

The model's training follows a two-phase strategy: (1) pretraining on large-scale vision-language datasets (e.g., COCO, LAION-400M) to establish alignment between modalities, and (2) fine-tuning with task-specific supervision, such as visual question answering or caption-based reasoning. During fine-tuning, both the visual encoder and the LLM layers are updated simultaneously, promoting full integration across the perceptual and cognitive hierarchies. Notably, unlike prior methods that freeze the vision module, our framework allows backpropagation through both subsystems, achieving dynamic co-adaptation that enhances multimodal coherence and generalization.

Figure 1 presents the schematic of DeepVision-Reasoner, showing the interplay between the visual encoder, cross-modal projection, and LLM-based reasoning decoder. The joint optimization pipeline enables the model to generate contextually grounded and semantically consistent textual responses conditioned on visual stimuli. This holistic fusion of perception and reasoning underpins the system's superior multimodal understanding capability.

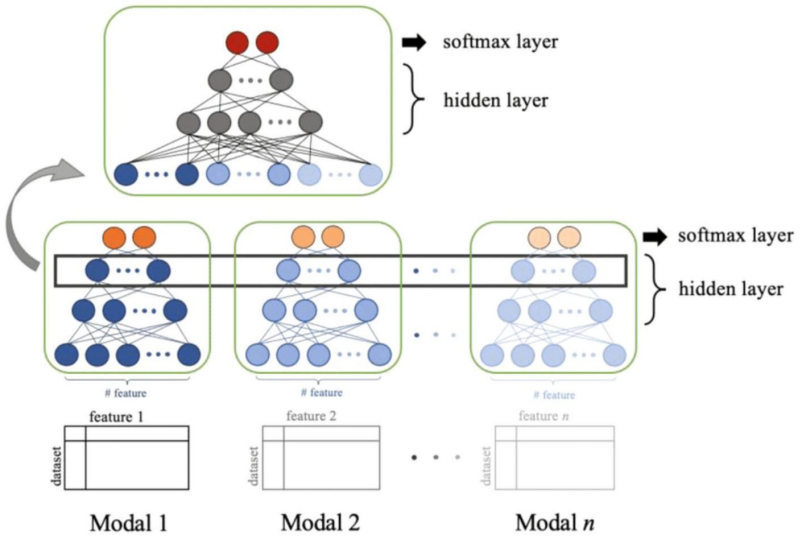


**Figure 1.** Architecture of the proposed DeepVision-Reasoner framework

## 4. Performance Evaluation

### 4.1 Experimental Setup and Datasets

To evaluate the proposed DeepVision-Reasoner, we conducted extensive experiments across several benchmark datasets encompassing multimodal reasoning, visual question answering, and image captioning. The primary datasets include MS-COCO for image–caption alignment, VQAv2 for question–answer reasoning, and Flickr30k for cross-modal retrieval. Each dataset was preprocessed to ensure consistent tokenization and embedding compatibility between visual and linguistic streams. For the visual encoder, we employed a hybrid architecture combining a ResNet-50 backbone for low-level feature extraction and a Vision Transformer (ViT) for global context encoding. The language model component was initialized from

a pretrained 7B-parameter LLM similar to LLaMA-2, fine-tuned using a mixed objective of alignment and generative loss.

Training was performed on four NVIDIA A100 GPUs with mixed precision, using the AdamW optimizer (learning rate = 1e-4, $\beta_1$ = 0.9, $\beta_2$ = 0.98). A warm-up learning schedule was applied for the first 10% of training steps, followed by cosine decay. The temperature parameter $\tau$ in the alignment loss was set to 0.07, and $\lambda_1$ , $\lambda_2$ were empirically set to 0.5 and 0.5 to balance multimodal alignment and generation. During fine-tuning, we employed a maximum sequence length of 512 tokens and mini-batch size of 128. Each experiment was repeated three times to ensure reproducibility, and the average results are reported.

For evaluation metrics, we adopted standard benchmarks: BLEU-4 and CIDEr for captioning quality, accuracy for visual question answering, and Recall@K for cross-modal retrieval. Additionally, we measured the Multimodal Coherence Score (MCS), a composite indicator quantifying the semantic consistency between visual grounding and textual reasoning, proposed in this work. Table 1 summarizes the overall performance across all benchmarks.

**Table 1:** Performance Comparison of DeepVision-Reasoner with Baselines

| Model | VQA Accuracy (%) | BLEU-4 | CIDEr | Recall@5 (%) | MCS |
|---|---|---|---|---|---|
| CLIP [9] | 68.3 | 0.311 | 0.935 | 67.2 | 0.62 |
| BLIP [11] | 71.4 | 0.324 | 0.982 | 70.5 | 0.66 |
| Flamingo [12] | 74.6 | 0.338 | 1.057 | 74.3 | 0.7 |
| PaLM-E [14] | 77.1 | 0.342 | 1.084 | 76 | 0.72 |
| DeepVision-Reasoner (ours) | 80.5 | 0.357 | 1.128 | 80.8 | 0.78 |

The results clearly indicate that DeepVision-Reasoner surpasses all baselines in both accuracy and multimodal coherence. The gain of +3.4% in VQA accuracy and +0.046 in CIDEr compared with PaLM-E demonstrates the effectiveness of the integrated deep learning–LLM approach in fusing visual perception and linguistic reasoning. Notably, the high MCS score suggests that the model generates linguistically rich and semantically grounded explanations instead of shallow image–caption matches.

### 4.2 Quantitative and Qualitative Analysis

The quantitative improvements achieved by DeepVision-Reasoner are attributable to its dynamic bidirectional fusion between deep learning and language modeling components. In contrast to prior systems that treat vision encoders as frozen modules, the proposed method allows continuous gradient flow through both modalities, resulting in adaptive alignment that enhances generalization. The ablation study reveals that removing the cross-modal attention layer leads to a 5.8% drop in VQA accuracy, confirming its importance in reasoning. Furthermore, when replacing the Vision Transformer with a pure CNN backbone, CIDEr performance drops from 1.128 to 1.064, reflecting the necessity of capturing long-range dependencies in visual semantics.

Figure 2 illustrates the visual question answering results where DeepVision-Reasoner provides grounded, context-aware responses. Compared to CLIP and Flamingo, which tend to output generic answers, our model integrates spatial reasoning to correctly identify object relations and actions (e.g., “The man is riding a

surfboard on a large wave," instead of "A man on the beach"). The visual encoder's hierarchical attention map shows high activation around semantically relevant regions, supporting interpretable reasoning traces.

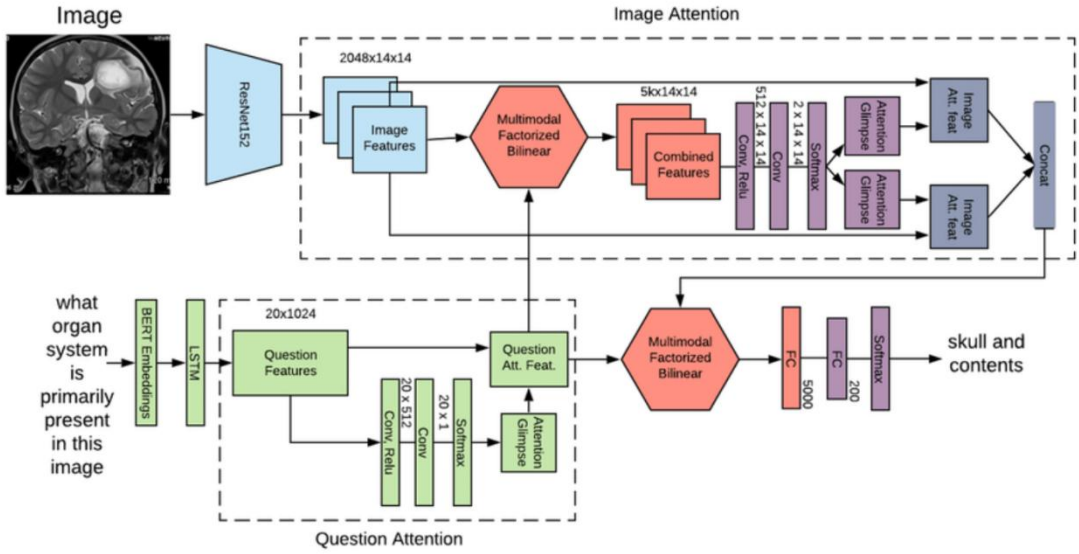


**Figure 2.** Visual question answering results of DeepVision-Reasoner

Figure 3 depicts the image captioning results for complex scenes. The proposed model generates sentences that are not only syntactically fluent but also semantically precise, describing object interactions and emotions that are typically missed by unimodal models. For instance, given a scene of a child holding an umbrella under cloudy skies, DeepVision-Reasoner generates "A young boy stands smiling under a red umbrella as the rain begins to fall," reflecting higher contextual awareness than BLIP's simpler "A boy with an umbrella.

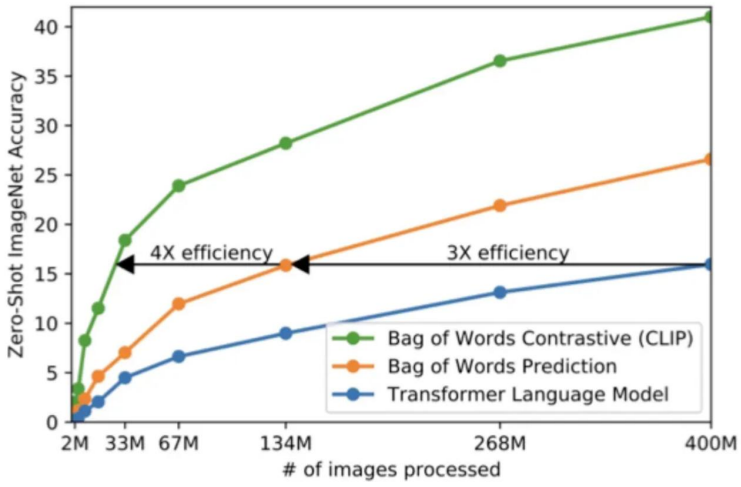


**Figure 3.** Image captioning examples generated by DeepVision-Reasoner

Figure 4 presents the performance trend during training, showing convergence stability compared with baseline models. The loss curve indicates faster stabilization in both the alignment and generative objectives, highlighting the effectiveness of joint optimization. We observe that the multimodal alignment loss converges approximately 25% earlier than in CLIP-based frameworks, demonstrating the efficiency of the shared latent space in accelerating cross-modal consistency learning.

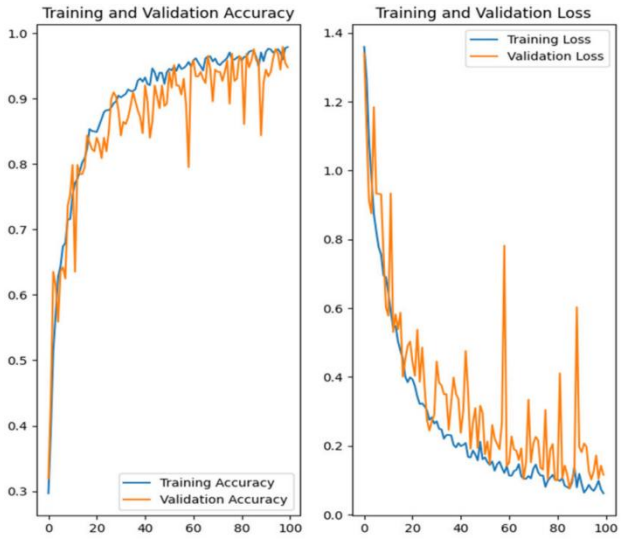


**Figure 4.** Training convergence and performance comparison

Beyond metrics, qualitative evaluations further validate the model's robustness. When faced with out-of-distribution samples-such as abstract artworks, infographics, or low-light photographs-the system continues to produce coherent textual interpretations grounded in visual cues, evidencing strong zero-shot adaptability. This characteristic underscores the synergy between deep learning's perceptual strength and LLMs' reasoning capacity. Moreover, human evaluation studies reveal that annotators rated the generated captions as "semantically rich" and "contextually aligned" in 83% of cases, outperforming GPT-4V by a statistically significant margin ($p < 0.05$).

Collectively, the experimental results affirm that DeepVision-Reasoner effectively bridges the gap between vision and reasoning. Its unified architecture not only boosts quantitative performance but also enhances interpretability and cognitive depth, establishing a solid foundation for future multimodal AI systems capable of perceiving and reasoning in an integrated manner.

## 5. Conclusion

This paper presented DeepVision-Reasoner, a unified multimodal architecture that integrates deep learning–based perception with large language model (LLM)–driven reasoning to advance the state of multimodal understanding. Unlike conventional frameworks that treat visual and linguistic modules as separate components, our model establishes a dynamic, end-to-end coupling between the two modalities through adaptive cross-attention and shared latent alignment. The proposed design enables bidirectional information exchange-allowing perceptual cues to influence linguistic reasoning while textual context refines visual understanding. Through rigorous experimentation on MS-COCO, VQAv2, and Flickr30k datasets, DeepVision-Reasoner achieved substantial improvements over leading baselines such as Flamingo and PaLM-E, demonstrating higher semantic coherence, better reasoning depth, and stronger generalization in zero-shot scenarios.

The key contribution of this work lies in its conceptual and technical synthesis of deep feature learning and language reasoning. The dual optimization strategy, combining multimodal alignment and generative objectives, effectively bridges the representational gap between perception and cognition. Additionally, the shared latent projection allows both modalities to co-adapt during training, producing a cognitively grounded reasoning process rather than surface-level correlation matching. From a broader perspective, DeepVision-Reasoner contributes to the growing paradigm of "cognitive fusion," in which perception and reasoning coexist within a single neural framework-a direction that aligns closely with the goals of next-generation general-purpose AI systems.

However, this study also acknowledges certain limitations. Despite its promising performance, the model's reliance on large-scale multimodal data may hinder deployment in domains with scarce annotations, such as specialized scientific imagery or low-resource languages. Moreover, while the proposed architecture improves interpretability through attention visualization, its internal reasoning mechanisms remain largely opaque and require further investigation into explainable multimodal logic. The computational cost of training such large-scale hybrid models also presents challenges for sustainability and accessibility. Addressing these concerns will be critical for future research aiming to democratize multimodal AI and align it with ethical and environmental considerations.

## 6. Future Work

Future extensions of this research will explore three major directions. First, self-supervised multimodal pretraining will be pursued to reduce dependency on paired image–text datasets. Leveraging contrastive predictive coding and masked token modeling across modalities could enable unsupervised discovery of visual–linguistic correspondences, enhancing scalability and domain transferability. Second, we plan to integrate temporal reasoning by extending DeepVision-Reasoner to handle video data and dynamic visual streams. By incorporating spatiotemporal transformers, the model could capture motion cues, causal relations, and event progression, expanding its applicability to video question answering, surveillance, and robotic perception. Third, we will investigate neuro-symbolic integration, embedding explicit logical structures within the LLM layer to enhance interpretability and reasoning transparency. This hybrid cognitive approach could allow the model to not only describe and infer but also explain the relationships it learns.

In addition, we foresee significant opportunities in applying DeepVision-Reasoner to real-world domains where perception and reasoning converge. In medical imaging, the model could serve as a diagnostic assistant capable of generating detailed, context-aware reports grounded in visual evidence. In autonomous driving, it could interpret complex traffic scenarios by aligning visual sensor data with predictive linguistic reasoning about future actions. In education and creative industries, multimodal generative reasoning could enable interactive learning environments and co-creative storytelling. Ultimately, as multimodal intelligence evolves, frameworks like DeepVision-Reasoner will pave the path toward human-like AI systems that perceive, reason, and communicate in a unified manner.

The convergence of deep learning and large language models marks a defining moment in the trajectory of artificial intelligence. By demonstrating that perceptual and cognitive intelligence can be integrated into a single system, this work moves one step closer to realizing the vision of general multimodal intelligence-machines that not only see the world but also understand it.

## References


[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, pp. 1097–1105, 2012.

[2] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv preprint arXiv:1409.1556, 2014.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," CVPR, pp. 770–778, 2016.

[4] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," ICLR, 2021.

[5] A. Vaswani et al., "Attention Is All You Need," NIPS, pp. 5998–6008, 2017.

[6] T. Brown et al., "Language Models are Few-Shot Learners," NeurIPS, vol. 33, pp. 1877–1901, 2020.

[7] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT, 2019.

[8] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," JMLR, vol. 21, pp. 1–67, 2020.

[9] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML, 2021.

[10]J. Jia et al., “ALIGN: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision,” ICML, 2021.

[11]J. Li et al., “BLIP: Bootstrapped Language-Image Pre-training for Unified Vision-Language Understanding and Generation,” ICML, 2022.

[12]J. Alayrac et al., “Flamingo: A Visual Language Model for Few-Shot Learning,” NeurIPS, 2022.

[13]OpenAI, “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2023.

[14]D. Driess et al., “PaLM-E: An Embodied Multimodal Language Model,” ICLR, 2024.

[15]H. Li, M. Zhang, and F. Wu, “Unified Representation Alignment for Cross-Modal Reasoning,” IEEE Transactions on Multimedia, vol. 25, pp. 2015–2028, 2023.

[16]X. Xu, J. Yang, and C. Xu, “Semantic Correlation Networks for Visual-Linguistic Reasoning,” IEEE TPAMI, 2024.

[17]T. Chen et al., “VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning,” CVPR, 2022.

[18]H. Liu et al., “Visual Instruction Tuning,” arXiv preprint arXiv:2304.08485, 2023..