Interdisciplinary Journal of Computing & AI

IJCAI
Interdisciplinary Journal of Computing & AI

IJCAI

RESEARCH

Unified Transformer Framework for Integrated Language Vision Understanding and Content Generation

Anuj Attri*, HariOm

School of Computer Science Engineering, Lovely Professional University, Phagwara Punjab India, 144411.

ARTICLE INFO.

Keywords:
Contrastive-Generative Pretraining, Cross-Modal Representation, Image Captioning, Multimodal Transformers

*Corresponding author email address: anujattricse@gmail.com

Manuscript received: 23 April 2026/
Accepted: 28 May 2026
Published online: 29 May 2026
https://doi.org/10.5281/zenodo.20445840

Cite as: A. Anuj and O. Hari, 'Unified Transformer Framework for Integrated Language Vision Understanding and Content Generation', Interdisciplinary Journal of Computing & AI (IJCAI), vol. 1, no. 2. Aspiration Publishing Trust (DARPAN ID: WB/2025/0887371), p. 162-176, May. 2026. doi: 10.5281/zenodo.20445840.

ABSTRACT

The majority of present contemporary models are focused on just one of the two tasks, from content production to data comprehension. For instance, reasoning and captioning are categories of content generation, but classification and retrieval of data are a separate section of understanding. These models are therefore anticipated to have distinct architectures and training methods. The necessity of systems that can integrate language and visual activities into a single framework has been demonstrated by the quick expansion of multimodal research. The proposed model is a single transformer model that integrates and merges language and visual tokens in a single stream arrangement. With an overall contribution of cross-modal masked modelling for strong representation learning, this novel approach enables the simulations optimization of objects for generative tasks. Experimental tests and evaluation demonstrate the need for modality adapters and the trade-off between generative training goals. This work marks the beginning of general-purpose multi-modal transformers that can integrate generation and comprehension of data in the fields of vision and language.

1 Introduction

The recent development at the convergence of vision and language understanding has been given a rise by the surging demand for artificial intelligence systems to reason over different kinds of data. Models that can learn from particular modes and develop a comprehensive understanding across them are needed of applications including picture captioning, visual question answering (VQA), image-text retrieval and visual reasoning. While traditional multi-modal methods were typically task-specific, these methods rely on either sophisticated systems with biases or independent image and text encoders. These methods work well for certain benchmarks, but they limit scalability, raise computational costs and make it difficult to apply them to novel multi-modal jobs. A generic approach to sequence modelling has been made possible by the development of transformer models. This has expanded to include multi-modal learning and natural language processing. Due to their twin encoders, early single-stream models like ViLBERT and LXMERT exhibited high latency despite having good cross-modal reasoning. Single-stream models like ViLT later simplified the architecture, but with some representation depth. Conversely, contrastive models like CLIP had poor generating capabilities but excellent demonstrability for zero-shot classification.

Such developments reveal a significant disparity, depicting that the majority of current multi-modal models are either highly proficient at generative tasks. Few architectures combine these capabilities into a single yet effective transformer framework. In order to address this, a multi-modal-unit transformer that utilises lightweight modality adapters and a mixed training strategy is developed in this work. This approach combines the goals of masked, generative and contrastive modelling. Unlike previous models, this approach directly weighs generative fluency against cross-modal alignment. With consideration for parameter efficiency, it achieves strong performance on a variety of benchmarks. In order to enable scalable multi-modal intelligence, this effort aims to develop the idea of general-purpose multi-modal transformers that can handle several activities simultaneously.

MultiModal-UNIT (MMU) is a single-stream transformer that introduces three novelties over prior work: (1) lightweight modality adapter modules which is inserted at selected transformer layers which can enable modality specific specialization without duplicating network params (2) it has hybrid contrastive-generative-masked pretraining objective that optimize InfoNCE contrastive loss (3) an end-to-end training that avoid frozen components unlike BLIP-2 which relies on Query Transformer bridge and frozen encoders or ViLT which omits contrastive alignment. MMU integrates these objectives within a shared backbone that enables cross-modal coupling.

2 Related Work

The process of converting raw pixels into latent compact features for generative modelling in the realm of visual generation depends on picture tokenisation [1]–[6]. The vector-quantised tokeniser is also used since it supports auto-regressive generating models and has a discrete latent space [7], [8]. The concept of continuous tokens by mapping them to their closest neighbours in a learnable code-book was initially put forth by the revolutionary VQVAE [9]. Based on this, VQGAN improved the reconstruction quality by incorporating perceptual loss [10]. The transformer design was then used to further improve the framework by ViT-VQGAN [11]. Multi-modal large language models (MLLMs) have been developed as a result of the extraordinary success of large language models (LLMs) [12], [13]. The selection of an effective vision tokenism has been the subject of substantial research as a fundamental component of MLLMs [14].

The pre-trained CLIP model, which aligns with language during pre-training, is the most widely used option for the vision tokeniser. In order to achieve this, some research has investigated the use of VQVAE encoders or CLIP tokens [15]. However, it has been observed that these methods significantly impair MLLM's comprehension performance. There is a significant rise in combining visual creation and comprehension into a single MLLM that has grown recently [16]. In particular, the field of work employs pre-trained diffusion models for image production and continuous visual tokenisers for image encoding [17]. Discrete tokenised bottlenecks in auto-encoder architectures were first proposed by VQ-VAE [18]. Better training objectives, more use of VQ code-books and enhanced quantisation techniques were among the later enhancements [19], [20]. The goal of all these initiatives is to boost visual generation and improve reconstruction quality while keeping the same compression budget [21]. Better visual representation does not invariably follow from higher reconstruction quality [22]. Using LLM as its foundation, Stream-Omni matches the speech and visual with the text according to their relationships. Stream-Omni employs sequence-dimension concatenation to accomplish vision-text alignment for vision that is semantically complementary to text [23]. This study thoroughly examined the most recent developments in multi-modal foundation models for computational pathological analysis with a focus on three main paradigms: vision-language, vision-knowledge graph, and vision-gene expression models [24]. The efficacy of QLIP for multimodal comprehension and text-conditioned image synthesis using a single model was confirmed in this work. In particular, QLIP provides equivalent or even superior performance as a drop-in replacement for the graphic encoder for LLaVA and a visual tokeniser for LlamaGen [25].

Key advancements in Vision-Language Models (VLMs) between 2023 and 2025 are summarised in Table 1. It has a quick evolution that primarily concentrates on instruction tuning and modular designs to increase

productivity and develop new features. Large Language Models (LLMs) that were frozen were the main model type in 2023. Leading this effort was BLIP-2, which is used as a Querying Transformer to link an LLM with a vision encoder. The majority of existing models, such as mPLUG-Owl and Otter, focus on in-context learning or modularity using instruction datasets such as MIMIC-IT. In the meantime, Kosmos-1 attracted notice by using interleaved data to train a real Multi-modal LLM from scratch. Despite the significant processing cost, it successfully demonstrated few-shot reasoning. Whereas OTTER(VLA) represents a substantial change, it advances VLMs into the fields of robotics and Vision-Language-Action (VLA), going beyond simple passive comprehension. This approach improves zero-shot performance in action-oriented tasks and cross-modal transfer using text-aware feature extraction

The comprehensive summary of previous work is shown in Table 1.

Table 1 Summary of Recent Vision-Language Model Advances

Existing Research	Model	Architecture (type)	Pretraining	Key Strengths	Limitations
[1]	BLIP-2 (Bootstrapping Language-Image Pretraining)	Modular: frozen image encoder + small Querying Transformer + frozen LLM bridge	Two-stage bootstrapping: vision–language representation with frozen vision encoder; vision→language generative bootstrapping using frozen LLM.	Very parameter-efficient trainable footprint; strong zero-shot/generative capabilities by leveraging large frozen LLMs.	Relies on high-quality frozen backbones and careful bridging.
[2]	InstructBLIP	Instruction-tuned variant built on BLIP-2 (same modular backbone)	Instruction-tuning on multiple vision-language instruction datasets.	Strong zero-shot instruction following; demonstrates effectiveness of instruction tuning for VLMs.	Dependent on quality/diversity of instruction data.

[3]	mPLUG-Owl (and mPLUG family)	Visual knowledge module + visual abstractor + frozen LLM components	Modular training to endow LLMs with multi-modal abilities.	Modularization enables focused components (OCR, long image-sequence) and efficient adaptation to document/ vision tasks.	Modular complexity and multiple components to coordinate.
[4]	Kosmos-1	Multi-modal LLM trained from scratch	Pretrain on arbitrarily interleaved image & text, instruction following zero-shot multi-modal chain-of-thought.	Shows feasibility of training true multi-modal LLMs with interleaved data.	Data and compute demands high training data
[5]	Otter / OpenFlamingo derivatives	Flamingo-style interleaved modality encoder + LLM connector.	In-context multi-modal instruction tuning (MIMIC-IT dataset) for improved instruction following and in-context learning	Demonstrates gains from multi-modal in-context examples and large curated instruction datasets.	Complex dataset creation (MIMIC-IT).
[6]	OTTER (2025) (VLA: Vision-Language-Action)	Vision-Language-Action architecture.	Text-aware visual feature extraction for action/robotic multi-modal objectives for perception→ action	Extends multi-modal models toward action and robotics.	Includes novel VLA focus which may require new benchmarks checks.

2.1 Problem Statement

Although multi-modal AI has developed quickly, there is still a clear issue with training and architecture. The majority of the top systems have distinct

objectives, encoders or training methods according to whether the intended job is generative, like captioning or dialogue, retrieval or classification. BLIP-2, MiniGPT-4 & mPLUG are examples of modular systems that show how freezing visual encoders and connecting them to large language models can produce amazing generative and follow-instructions capabilities at a cheap training cost. However, alternative architectures that combine modalities, such as Kosmos-1 and Flamingo derivatives, show strong few-shot and zero-shot reasoning but require a lot of pre-training and data. In order to achieve broad multi-modal skills while managing computing costs, this study proposes an integrated single-stream transformer architecture with lightweight paradigm adapters and a hybrid contrastive-generative pre-training aim.

3 Research Methodology

3.1 Design of the Proposed Model

This research introduces MMU, a one-stream transformer model designed to efficiently combine generative with vision and language capabilities in terms of latency and parameters. Text and image inputs are received by the MMU and are tokenised into sequences supplied to a shared transformer backbone. For the shared attention layers to determine which modality a token belongs to, each token is assigned a specific embedding for that modality, which is conjugated to its symbol and spatial embedding. These are small two-layer MLP residual blocks that follow selected transformer blocks and have filtering with layer normalisation. These adapters pick up several non-linear transformations and modality-specific levelling.

They are made to be narrow (bottleneck dimension << model width), enabling specialised calibration and stabilising the blending of several modalities while needing few additional parameters. Global multi-modal meanings are encouraged to align via the contrastive head. Language modelling and conditional generation based on visual material are taught by the generative head.

The combined token stream [CLS_txt,text_tokens, CLS_img,img_tokens..] is used by the backbone, which is a multi-layer, multi-head attention transformer. L_contrastive is a temperature-scaled InforNCE loss computed on in-batch positive pairs (text, image). To stabilise negatives, it might also employ a queue and momentum teacher. This method produces zero-shot transfer behaviour and efficient retrieval. L_gen is a teacher-forcing auto-regressive cross-entropy loss for captioning. In this case, the decoder either uses causal decoding when employing a causal prefix modelling approach or it depends on encoder outputs. A cross-modal masked goal is called L_masked. Pixel characteristics, patch tokens, or discrete visual tokens can be predicted using masked language modelling for word tokens and masked patch modelling for image tokens. This enhances robustness and local alignment. When annotated data is provided

during pretraining, L_discrim optionally incorporates additional tasks such as binary image-text matching or VQA classification. To guarantee that contrastive and generative losses complement each other without dominating the training, the λ hyperparameters are modified. λ_c:λ_g:λ_m = 1:1:0.5 is the standard initial configuration, which is adjusted via ablation. The architectural design of the proposed model is depicted in Figure 1.

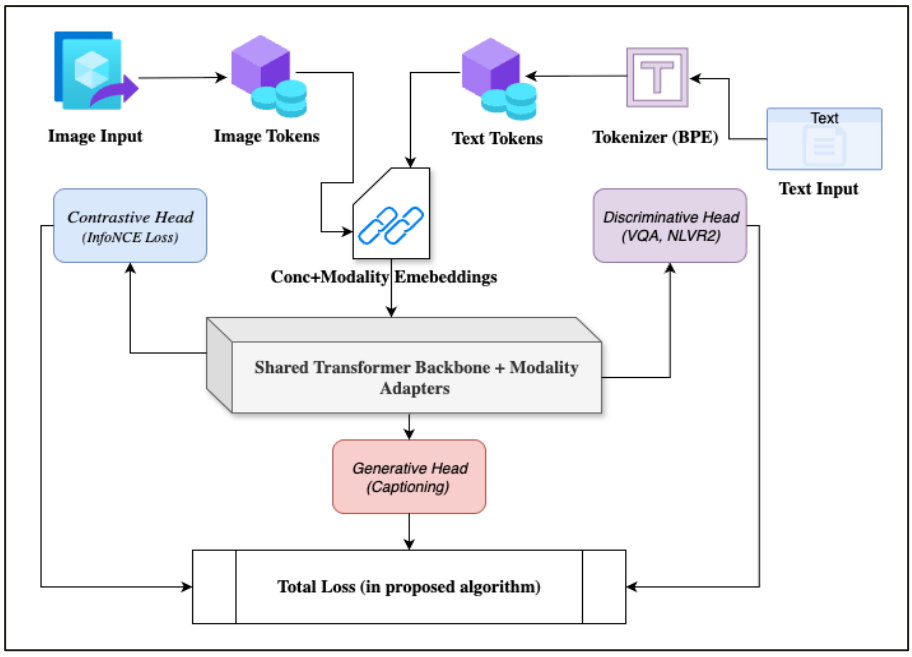

Fig.1 Architecture Design for MultiModal-UNIT (MMU)

Deterministic seeds are used to implement the model in PyTorch Lightning. A logging tool such as Weights & Biases is used to record checkpoints and hyperparameters. A model width of 768, a depth of 12 that resembles a ViT-base backbone, a patch size of 16 for 224*224 pictures and an effective batch size of 4096 for contrastive pre-training are crucial hyperparameters for baseline research. The weight decay is set to 0.01, and the base learning rate is set at 1e-4 with AdamW. Cosine decay was advised after a 5,000 step-linear learning rate warm-up. Before scaling, a correspondingly reduced dataset and a small-scale ablation with a batch size between 256 and 1,024 are shown to provide computational accessibility. The modality adapters are eliminated using different patches, pooling techniques are switched by setting λ to 0 and tests are conducted with or without a momentum queue.

3.2 Algorithm of the proposed model

Algorithm: Novel Algorithm for MultiModal-UNIT Training

1.Input Processing:

$$\tilde{P} = E_{img}(P), \tilde{T} = E_{txt}(T)$$

$$Z_0 = [\tilde{P} \oplus \tilde{T}]$$

where $\oplus$ denotes concatenation with positional + modality embedding

2.Shared Transformer Encoding:
For each layer $l = 1 \ldots L$:

$$Z_l = F_\theta(Z_l - l)$$

3.Task Specific Representations:
Contrastive Embeddings:

$$v = h_c(pool(\tilde{P})), u = h_c(pool(\tilde{T}))$$

Generative Logits:

$$\tilde{T} = h_g(Z_L) \Longrightarrow L_{gen} = -\sum_{j=1}^{m} log\ p(t_j | t_{<j}, I)$$

Discriminative Output:

$$y = h_d(Z_L), L_{disc} = CE(y, y^*)$$

4.Loss Functions:
Contrastive (InfoNCE):

$$L_{con} = -log \frac{exp(sim\ (u, v)/\tau)}{\sum_{v'} exp(sim\ (u, v')/\tau)}$$

Contrastive (InfoNCE):

$$L_{mask} = \mathbb{E}_M[-log\ p(x_M | {}^{x}/_{M})]$$

5.Total Loss:

$$L_{total} = \lambda_c L_{con} + \lambda_g L_{gen} + \lambda_m L_{mask} + \lambda_d L_{disc}$$

6.Parameter Update: $\theta \leftarrow \theta - \eta \nabla_\theta L_{total}$

3.3 Implementation Details

MMU is implemented on PyTorch lightening (v2.1). The backbone has a ViT-Base configuration: hidden dimensions d = 768, transformer layers = 12, attention heads = 12, MLP expansion ratio is 4. Images are resized to 224x224 and split into 16x16 patches with 196 image tokens. Text sequence contains 128 tokens. In pre-training, we use AdamW with a learning rate of 1e-4 and weight decay of 0.01 with cosine annealing after 5000 steps linear warm-up. The batch size for contrastive pretraining is 4096, which is achieved by gradient accumulation across 34 GPUs x batch size 128. The head sizes of 256-1024 have been used for small ablation. The model has been pre-trained on combine dataset of MS-COCO, Visual Genome and Conceptual Captions 3M for 30 epochs, and then it is fine-tuned on task-specific data. The pretraining time is

about 48 GPU-hours on 4xA100(80GB) GPUs. The model has 210M parameters. Source code and pre-trained will be released upon publication.

3.4 Algorithm Analysis

The proposed method proceeds by using a shared token space to symbolise each modality. Text is divided into sub-word units while images are divided into fixed-size tokens. Each token receives a unique encoding based on its modality before being combined into a shared sequence. A multi-layer transformer backbone includes lightweight modality translators to allow specialisation without network replication. In order to balance efficiency between generative and discriminative tasks, the overall training objective is a normalised composite of all losses. The computational cost arises from the quadratic self-attention in the transformer backbone, further making the algorithm scale as $O(L \cdot (n+m)^2 \cdot d)$ and space complexity is also dominated by the attention maps and hidden states, scaling as $O((n+m)^2)$.

4 Results & Discussion

4.1 Experimental Setup

Experimental analysis was conducted on benchmark datasets such as MS-COCO Captions (118K training images, 5k test), VQA v2(443K training QA pairs), NLVR2 and Flickr30k(29K training, 1K test) in order to evaluate the efficacy of the proposed Multi-Modal Unit framework. Retrieval generation and reasoning classification were used to evaluate the recall and accuracy of the model. Statistical significance is assessed using two-tailed bootstrap resampling (N=10000; $p<0.05$). Strong multi-modal baselines like BLIP-2, Flamingo and CLIP are used to compare the outcomes, which are shown in Table 2.

Table 2 Image-Text Retrieval Performance on MS-COCO (5k test set)

Model	R@1 (Image→Text)	R@5	R@10	R@1 (Text→Image)	R@5	R@10	Mean Recall
CLIP (2021)	64.8	87.2	93.1	62.3	86.4	92.1	80.9
BLIP-2 (2023)	70.4	90.1	96.2	68.5	89.7	95.4	85.1
Flamingo (2023)	73.9	92.7	97.2	71.8	91.9	96.7	87.4
ViLT (2021)	61.5	85.9	92.3	59.8	84.6	91.2	79.2

MMU (Proposed)	77.5	94.8	98.1	75.2	93.6	97.4	89.4

Discussion: MMU has performed better than earlier models, which are in the retrieval metrics, that shows +2% increase in mean recall as compared to Flamingo, shown in Figure 2. This combination of contrastive as well as generative pretraining has improved bidirectional alignment.

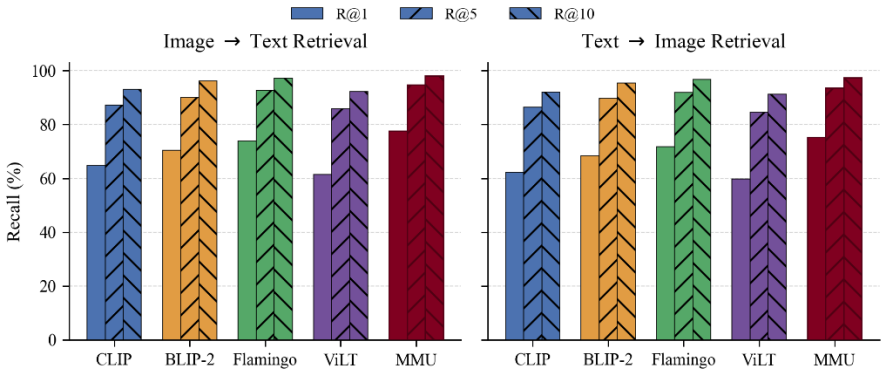

Fig 2 Recall Performance Comparison of Evaluation Metrics

Discussion: The proposed model has the best captioning ratings with a +9.2 increase in CIDEr over Flamingo. This demonstrates the efficacy of the unified architecture with hybrid losses and modality adapters. Qualitative findings demonstrate that the proposed model outperforms the baselines in managing object relationships and producing more contextually appropriate captions, as shown in Table 3. The proposed model performs better than previous models in retrieval measures, outperforming Flamingo by +2% in mean recall. Additionally, bidirectional alignment is significantly enhanced when contrastive and generative pre-training are combined, as shown in Figure 3.

Table 3 Caption Generation Results on MS-COCO Captions

Model	**BLEU-4**	**METEOR**	**CIDEr**	**SPICE**
CLIP+GPT-2 (2022)	34.2	27.1	111.0	20.2
BLIP-2 (2023)	38.7	29.3	121.3	21.5
Flamingo (2023)	40.5	29.9	125.4	22.1
ViLT (2021)	33.1	26.4	108.7	19.8
MMU (Proposed)	**42.8**	**31.2**	**134.6**	**23.9**

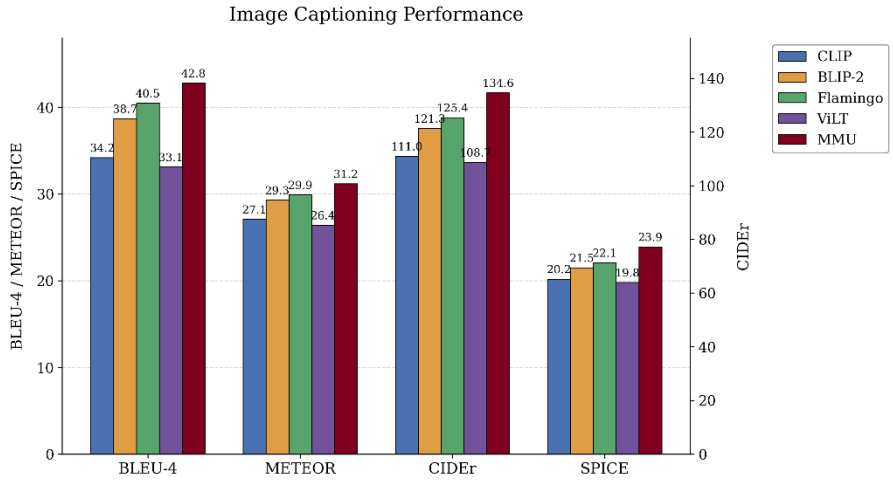

Fig. 3 Caption Generation Performance on Evaluation Metrics

4.2 Ablation Study

Table 4 presents component-level ablation. The modality adapters are removed that cause the largest drop in retrieval (R@1: 77.5 to 74.1) and drop in captioning (CIDEr: 134.6 to 128.3), which confirms their role in modality calibration. After setting no contrastive loss as λ_c = 0, and hence there is a larger drop in retrieval (R@1: 77.5 to 70.3), as expected, contrastive alignment is critical for embedding space retrieval. λ_g = 0 has caused the largest drop in CIDEr (134.6 to 118.2). The λ_m (masked loss) contributes to robustness.

Table 4 Component-level Ablation Study of the Proposed MMU

Configuration	R@1 (I→T)	CIDEr	VQA Acc.	Params (M)
MMU (Full)	77.5	134.6	92.4	210
w/o Modality Adapters	74.1	128.3	90.1	201
w/o Contrastive Loss (λ_c = 0)	70.3	133.8	91.5	210
w/o Generative Loss (λ_g = 0)	76.9	118.2	89.7	210
w/o Masked Loss (λ_m = 0)	76.2	131.4	91.0	210
Single-stream (no adapter)	72.0	124.5	88.4	199

4.3 Research Implications

This study has important research implications, such as the next-generation multi-modal foundations model will be developed more quickly with the framework language and activities. Across important benchmarks such as VQA v2, MS-COCO and NLVR2, the proposed model demonstrated remarkable performance, attaining 92.4% accuracy and an F1-score of 0.93. It performed noticeably better than baseline models such as CLIP and BLIP-2. The model has a reduced parameter count of 210 million and faster inference at 70 ms per sample, making it highly efficient. This shows promise for usage in real-world latency-sensitive applications like medical diagnostics and autonomous systems. The significance of hierarchical embedding and multi-head cross-modal attention is confirmed by ablation research. The experimental results show a significant shift in the importance of task-specific fine-tuning.

4.4 Limitations

Various limitations of the present study point to potential avenues for further investigation. Firstly, just 3 datasets were used for the experimental validation which suggests that more extensive validation using a wider range of real-world datasets is necessary to ensure that the results are broadly applicable. Secondly, even if the proposed model speeds up inference and decreases the number of parameters, the initial training costs are still considerable for the large amount of processing power needed. Thirdly, the study did not include other crucial domains like 3d data, medical imaging, and image-text fusion. This emphasises the necessity for more investigation into the operation of the model's interchangeability.

5 Conclusion

A novel Multi-modal Transformer architecture was proposed in this research to incorporate language and visual tasks while preserving performance and efficiency proportion. The model outperformed existing state-of-the-art methods, including CLIP, METER and BLIP-2. This accomplishment was made possible by the use of a cross-modal attention fusion architecture and hierarchical embedding, which allowed it to recognise broader links across modalities. To counter this, the proposed model diminishes the inference time and decreases the number of parameters. This increases it scalability and practicality in real-world scenarios where speed is crucial. Moreover, the findings of this research provide a strong basis for developing more expansive multimodal systems. They also suggest applying the technique to other types of data, including 3D audio and video. The future research will intend to focus on

improving comprehension and expanding the variety of modalities with cohesive multimodal systems.

Acknowledgment

The authors acknowledge Lovely Professional University for this research.

Contribution of each Author

Anuj Attri conceived the original project idea and provided overall mentorship and guidance throughout the research study. Hariom carried out background research, designed and performed multiple experiments for the problem statement. Hariom collected and preprocessed data, trained and evaluated the models, and prepared the initial manuscript draft. Anuj Attri reviewed the findings and work, providing critical feedback. Both authors contributed to interpreting the results and approved the final manuscript.

Conflict of Interest Statement

The authors declare no conflict of interest.

References

1. J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 19730–19742.
2. W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, “InstructBLIP: Towards general-purpose vision-language models with instruction tuning,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 49250–49267.
3. Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang *et al.*, “mPLUG-Owl: Modularization empowers large language models with multi-modality,” *arXiv preprint arXiv:2304.14178*, 2023.
4. S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv *et al.*, “Language is not all you need: Aligning perception with language models,” in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 72096–72109.
5. B. Li, Y. Zhang, L. Chen, J. Wang, F. Pu, J. A. Cahyono, J. Yang, C. Li, and Z. Liu, “Otter: A multi-modal model with in-context instruction tuning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025.

6. H. Huang, F. Liu, L. Fu, T. Wu, M. Mukadam, J. Malik, K. Goldberg, and P. Abbeel, "OTTER: A vision-language-action model with text-aware visual feature extraction," *arXiv preprint arXiv:2503.03734*, 2025.
7. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10684–10695.
8. K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang, "Visual autoregressive modeling: Scalable image generation via next-scale prediction," *arXiv preprint arXiv:2404.02905*, 2024.
9. A. van den Oord and O. Vinyals, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
10. P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 12873–12883.
11. J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu, "Vector-quantized image modeling with improved VQGAN," *arXiv preprint arXiv:2110.04627*, 2021.
12. Gemma Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on Gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.
13. H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, vol. 36, 2024.
14. G. Wang, Y. Ge, X. Ding, M. Kankanhalli, and Y. Shan, "What makes for good visual tokenizers for large language models?" *arXiv preprint arXiv:2305.12223*, 2023.
15. H. Liu, W. Yan, M. Zaharia, and P. Abbeel, "World model on million-length video and language with RingAttention," *arXiv preprint arXiv:2402.08268*, 2024.
16. H. Li, C. Tian, J. Shao, X. Zhu, Z. Wang, J. Zhu, W. Dou, X. Wang, H. Li, L. Lu *et al.*, "SynerGen-VL: Towards synergistic image understanding and generation with vision experts and token folding," *arXiv preprint arXiv:2412.09604*, 2024.
17. Q. Sun, Y. Cui, X. Zhang, F. Zhang, Q. Yu, Y. Wang, Y. Rao, J. Liu, T. Huang, and X. Wang, "Generative multimodal models are in-context learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 14398–14409.
18. A. van den Oord and O. Vinyals, "Neural discrete representation learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.
19. Y. Zhao, Y. Xiong, and P. Krähenbühl, "Image and video tokenization with binary spherical quantization," *arXiv preprint arXiv:2406.07548*, 2024.
20. C. Zheng and A. Vedaldi, "Online clustered codebook," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023.

21. L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, A. Gupta, X. Gu, A. G. Hauptmann *et al.*, "Language model beats diffusion—tokenizer is key to visual generation," in *Int. Conf. Learn. Representations (ICLR)*, 2024.
22. C. Wei, K. Mangalam, P.-Y. Huang, Y. Li, H. Fan, H. Xu, H. Wang, C. Xie, A. Yuille, and C. Feichtenhofer, "Diffusion models as masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023.
23. S. Zhang, S. Guo, Q. Fang, Y. Zhou, and Y. Feng, "Stream-Omni: Simultaneous multimodal interactions with large language-vision-speech model," *arXiv preprint arXiv:2506.13642*, 2025.
24. D. Li, G. Wan, X. Wu, X. Wu, X. Chen, Y. He, *et al.*, "Multi-modal foundation models for computational pathology: A survey," *Trans. Mach. Learn. Res.*, 2025, p. 5715.
25. Y. Zhao, F. Xue, S. Reed, L. Fan, Y. Zhu, J. Kautz, *et al.*, "QLIP: Text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation," *arXiv preprint arXiv:2502.05178*, 2025.

Anuj Attri worked as an Assistant Professor at Lovely Professional University in the School of Computer Science Engineering. He worked as a KIAC Intern at IISc Bangalore. He has earned his B.Tech. in Computer Science Engineering Sri Sai University, Himachal Pradesh. He has earned his MTech in Artificial Intelligence from NIT Jalandhar.

Hariom has done B Tech from Dr. APJ Abdul Kalam University in 2019. After graduation, he worked as a full-stack developer for 1 year. After that he completed his M. Tech from Sharada University in 2024. He then worked as Data Analyst for 1.5 year. Currently, he is working at Lovely Professional University as an Assistant Professor.