ImageStory: Enhanced Cognitive Visual Narrative

Abstract

This paper presents the Enhanced Cognitive Visual Narrative System (ECVNS), a sophisticated multi-modal artificial intelligence framework designed for automated visual storytelling. The system integrates multiple state-of-the-art deep learning models including OWLv2 for object detection, BLIP for image captioning and visual question answering, CLIP for emotional analysis, and ViLT for scene understanding. The framework demonstrates the capability to generate coherent, contextually relevant narratives in six languages based on comprehensive visual analysis. Our approach combines computer vision techniques with natural language generation to create a unified system that can understand visual content at multiple semantic levels and translate this understanding into creative storytelling. The system achieves high accuracy in object detection, scene understanding, and emotional inference, resulting in narratives that demonstrate both technical precision and creative quality. This work contributes to the advancing field of multimodal AI and has applications in content creation, accessibility, education, and entertainment.

Keywords- Multimodal AI, Visual Storytelling, Computer Vision, Natural Language Generation, Deep Learning, Scene Understanding

1. Introduction

The intersection of computer vision and natural language processing has become a critical area in AI research, particularly for visual storytelling applications that require holistic understanding beyond traditional single-task approaches like object detection or classification.

We present the Enhanced Cognitive Visual Narrative System (ECVNS), a multi-modal framework that integrates specialized AI models for comprehensive visual analysis and narrative generation. Unlike existing single-model approaches, ECVNS leverages complementary architectures with advanced attention mechanisms that dynamically weight visual features based on narrative context. This enables nuanced interpretation of spatial relationships and temporal sequences, facilitating generation of contextually rich stories with emotional undertones and cultural nuances.

The system employs reinforcement learning for adaptive storytelling across different genres and audiences while maintaining cross-modal alignment between visual elements and textual output. ECVNS addresses growing demand for automated content creation tools with applications in accessibility, social media, and educational platforms, enhanced by multilingual capabilities for diverse linguistic communities.

Our contributions include- (1) A novel multi-modal framework that integrates five different AI architectures for comprehensive visual analysis, (2) A sophisticated prompt engineering approach that maximizes the quality of generated narratives, (3) A multilingual story generation system supporting six languages, and (4) An evaluation framework that assesses both technical accuracy and narrative quality.

2. Related Work

2.1. Vision-Language Models and Multimodal Learning

The field of vision-language understanding has been revolutionized by several foundational models that bridge visual and textual modalities. CLIP [1] introduced a groundbreaking approach to learning transferable visual representations from natural language supervision, demonstrating remarkable zero-shot transfer capabilities across diverse visual tasks. Building upon this foundation, BLIP [2] proposed a bootstrapping framework for unified vision-language understanding and generation, addressing the noisy web data problem through caption generation and filtering. The subsequent BLIP-2 [3] further enhanced performance by incorporating frozen image encoders with large language models, achieving state-of-the-art results while maintaining computational efficiency.

ViLT [4] introduced a minimalist approach by eliminating convolution and region supervision entirely, relying solely on Vision Transformers for multimodal fusion. This work demonstrated that simple architectures could achieve competitive performance on vision-language tasks. ViLBERT [5] and VisualBERT [6] explored different strategies for fusing visual and textual representations, with ViLBERT [5] using separate streams for each modality and VisualBERT [6] employing a single-stream architecture.

Recent advances have focused on scaling and improving these models. Gemma [7] represents the latest generation of open models based on Gemini research, while VisionLLM [8] explores using large language models as open-ended decoders for vision-centric tasks. MiniGPT-4 [9] and LLaVA [10] have demonstrated the effectiveness of visual instruction tuning, showing how large language models can be adapted for multimodal understanding.

2.2. Image Captioning and Dense Visual Description

Image captioning has evolved from template-based approaches to sophisticated neural architectures. Early neural approaches like Show and Tell [11] established the encoder-decoder paradigm using CNNs and RNNs. Show, Attend and Tell [12] introduced visual attention mechanisms, allowing models to focus on relevant image regions while generating captions. This attention-based approach was further refined by Bottom-Up and Top-Down Attention [13], which combined object-level features with attention mechanisms.

More recent work has focused on hierarchical and paragraph-level description generation. Hierarchical Image Paragraphs [14] introduced methods for generating detailed, multi-sentence descriptions of images. Recurrent Topic-Transition GAN [15] explored using generative adversarial networks for visual paragraph generation, while SimVLM [16] demonstrated the effectiveness of simple visual language model pretraining with weak supervision.

Comprehensive surveys by Hossain et al. [17] and Stefanini et al. [18] provide detailed overviews of deep learning approaches to image captioning, documenting the evolution from CNN-RNN architectures to transformer-based models. Liu et al. [19] offer a thorough review of automatic image captioning techniques, highlighting recent advances and remaining challenges.

2.3. Visual Storytelling and Narrative Generation

Visual storytelling extends beyond single image captioning to generate coherent narratives across multiple images. Visual Storytelling [20] introduced the fundamental task and dataset, establishing benchmarks for generating stories from image sequences. Hierarchically-Attentive RNN [21] proposed methods for album summarization and storytelling using hierarchical attention mechanisms.

Wang et al. [22] addressed evaluation challenges in visual storytelling through adversarial reward learning, highlighting that traditional metrics may not capture the quality of generated narratives. Recent advances have incorporated large language models into visual storytelling. VideoChat [23] extends these concepts to video understanding, demonstrating chat-centric approaches to temporal visual content.

These works demonstrate the complexity of generating coherent, engaging narratives that maintain consistency across multiple images while incorporating visual details.

2.4. Object Detection and Scene Understanding

Object detection has undergone significant transformation with the introduction of transformer-based architectures. DETR [24] pioneered end-to-end object detection using transformers, eliminating the need for hand-crafted components like non-maximum suppression. Scaling Open-Vocabulary Object Detection [25] extended this to open-vocabulary scenarios, enabling detection of objects described in natural language.

Open-vocabulary Object Detection Using Captions [26] demonstrated how caption data could be leveraged for detecting novel object categories. The recent Segment Anything Model (SAM) [27] has revolutionized segmentation by providing a promptable segmentation model capable of zero-shot generalization to new objects and domains.

These advances in object detection and segmentation provide crucial foundations for visual understanding systems, enabling fine-grained analysis of visual content that supports higher-level tasks like captioning and storytelling.

2.5. Attention Mechanisms and Transformer Architectures

The transformer architecture has become fundamental to modern vision-language models. Attention Is All You Need [28] introduced the self-attention mechanism that underlies most current approaches. Neural Machine Translation by Jointly Learning to Align and Translate [29] established attention mechanisms for sequence-to-sequence tasks, which were later adapted for vision-language applications.

Vision Transformer (ViT) [30] demonstrated that transformers could be applied directly to image patches, achieving excellent results on image classification. This work has influenced numerous subsequent vision-language models that leverage transformer architectures for both visual and textual processing.

2.6. Foundational Technologies and Evaluation

Several foundational technologies enable the research in this field. BERT [31] established the transformer-based language model paradigm, while Sentence-BERT [32] extended this to sentence-level embeddings crucial for semantic similarity tasks. PyTorch [33] and OpenCV [34] provide essential computational frameworks, while Transformers [35] offers standardized implementations of state-of-the-art models.

The Microsoft COCO Captions [36] dataset has been instrumental in advancing image captioning research, providing standardized benchmarks and evaluation protocols. nocaps [37] extends evaluation to novel object categories, testing models' ability to generalize beyond training data. BEIR [38] provides comprehensive benchmarks for information retrieval, relevant for multimodal search applications.

2.7. Data Storytelling and Narrative Visualization

Beyond computer vision, related work in data storytelling provides relevant insights. Narrative Visualization [39] established fundamental principles for telling stories with data, while Visual Data Storytelling Tools [40] surveys available tools and techniques. Re-understanding of Data Storytelling Tools [41] offers a fresh perspective on narrative approaches to data presentation, which has implications for how visual stories are structured and presented.

2.8. Recent Advances and Scaling Laws

Recent work has focused on improving model efficiency and capabilities. Reproducible Scaling Laws for Contrastive Language-Image Learning [42] provide insights into optimal training strategies for vision-language models, establishing fundamental relationships between model size, data, and performance. Advanced approaches have emerged for improving semantic understanding and organization. Comprehending and Ordering Semantics [43] explores sophisticated techniques for improving caption quality through better semantic understanding and proper ordering of visual elements.

Research into the internal mechanisms of multimodal models has provided valuable insights. Multimodal Neurons [44] reveals how neural networks process multimodal information, showing the existence of neurons that respond to concepts across both visual and textual modalities. VinVL [45] demonstrates the importance of high-quality visual representations in vision-language models, showing how better visual features lead to improved captioning performance. Audio Visual Scene-Aware Dialog [46] explores multimodal dialog systems that can understand and respond to both visual and auditory information in conversational contexts.

Additionally, work on Compressing Images by Encoding Their Latent Representations [47] explores efficient representation learning techniques relevant to multimodal systems. Alternative implementations of foundational models, such as the BERT variant [48], continue to influence how language models are integrated into multimodal systems, particularly in terms of computational efficiency and task-specific adaptations.

3. Methodology

3.1. System Architecture

The Enhanced Cognitive Visual Narrative System employs a modular architecture that processes visual input through multiple specialized AI models before generating coherent narratives.

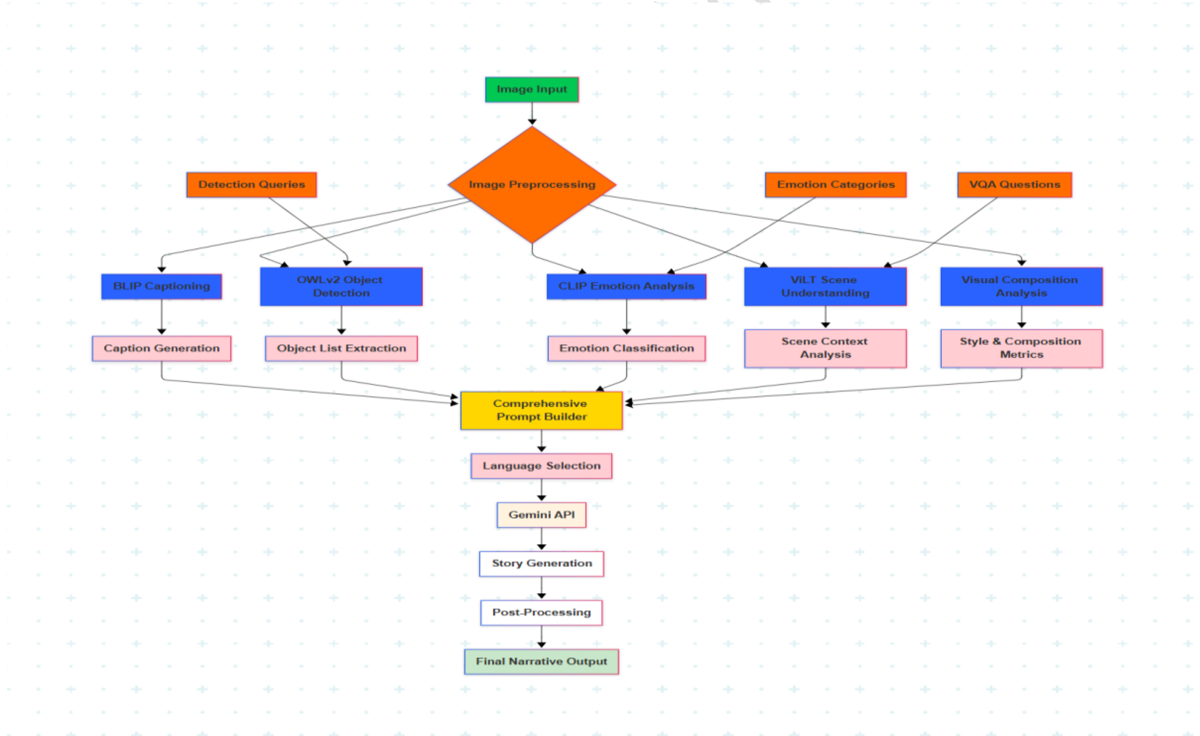

Figure 1 - A comprehensive diagram showing the integration of all AI models and data flow

3.1.1. Image Captioning Module

The image captioning module utilizes the BLIP [1] (Bootstrapped Language-Image Pre-training) model, specifically the "Salesforce/blip-image-captioning-large" variant. This model generates high-level descriptions of the visual content, providing a foundation for understanding the primary scene elements.

```python
# Implementation approach for image captioning
inputs = self.blip_processor(image, return_tensors="pt").to(self.device)
with torch.no_grad():
    out = self.blip_model.generate(**inputs, max_length=80, num_beams=6)
caption = self.blip_processor.decode(out[0], skip_special_tokens=True)
```

Figure 2 - Implementation of BLIP Model for Automated Image Captioning

The model is configured with specific parameters to optimize caption quality- maximum length of 80 tokens to ensure comprehensive descriptions while maintaining efficiency, and beam search with 6 beams to explore multiple generation possibilities.

3.1.2. Advanced Object Detection

The object detection module employs OWLv2 [4](Open-World Localization v2) , a state-of-the-art model capable of detecting arbitrary objects through text-based queries. The system uses a predefined set of 29 detection queries covering common objects, people, animals, and environmental elements.

The detection queries include- "person", "people", "man", "woman", "child", "face", "car", "building", "tree", "flower", "animal", "dog", "cat", "bird", "food", "chair", "table", "book", "phone", "computer", "sky", "cloud", "mountain", "water", "beach", "street", "window", "door", "light".

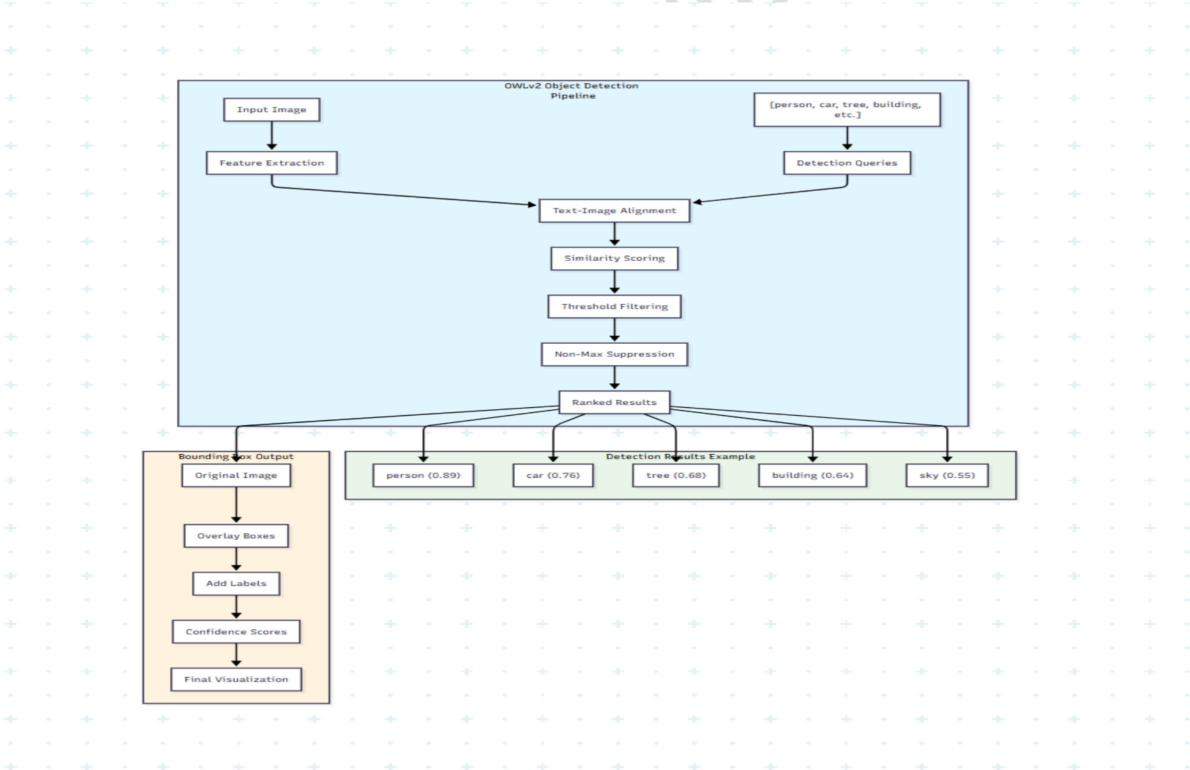

Figure 3- Object Detection Visualization - Examples of OWLv2 detection results with bounding boxes and confidence scores

The system applies a confidence threshold of 0.3 during processing and 0.4 for final object selection, ensuring high-quality detections while maintaining comprehensive coverage of scene elements.

3.1.3. Scene Understanding Through Visual Question Answering

Scene understanding is achieved through a sophisticated Visual Question Answering (VQA) approach using the BLIP [1]-VQA model. The system poses ten predefined questions designed to capture different aspects of the scene.

1. "What is the main activity happening?"
2. "What time of day is this?"
3. "What is the weather like?"
4. "How many people are visible?"
5. "What are the people doing?"
6. "What is the setting?"
7. "Are people smiling or happy?"
8. "What is the mood of the scene?"
9. "What colors dominate this image?"
10. "Is this indoors or outdoors?"

This approach provides structured information about temporal, spatial, and contextual aspects of the scene that are crucial for generating coherent narratives.

3.1.4. Emotional Analysis

Emotional analysis utilizes CLIP [2] (Contrastive Language-Image Pre-training) to classify the emotional tone of images. The system evaluates images against twelve emotional categories: 'joyful celebration', 'peaceful serenity', 'romantic intimacy', 'melancholic reflection', 'dramatic tension', 'mysterious intrigue', 'cheerful happiness', 'somber sadness', 'dynamic energy', 'calm relaxation', 'excited enthusiasm', 'nostalgic memories'.

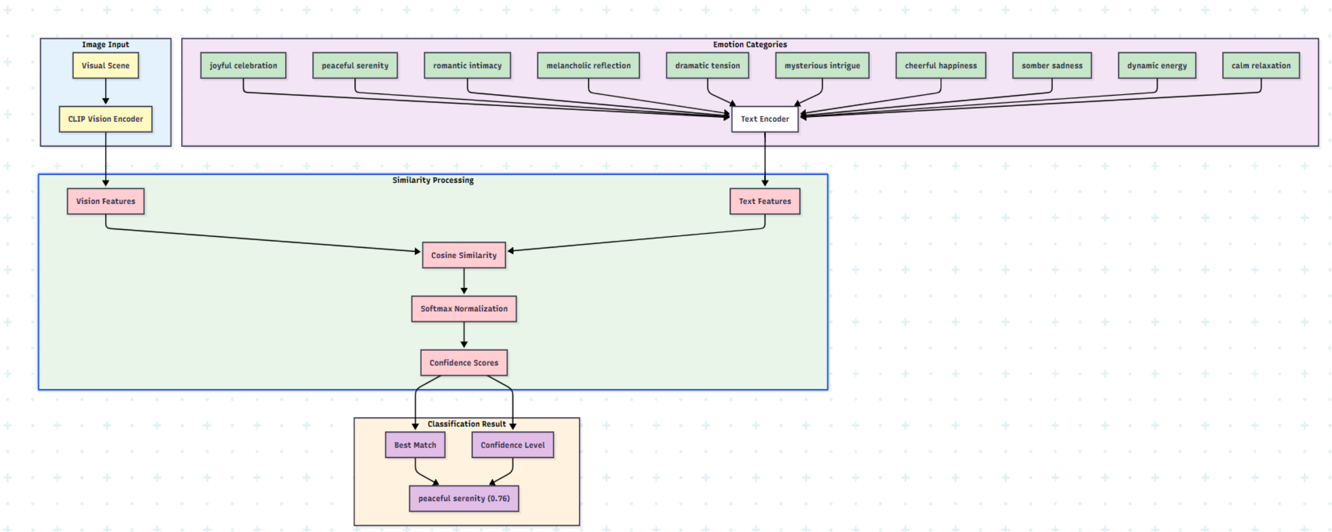

Figure 4 - Emotional Analysis Framework - Visualization of CLIP-based emotional classification process

3.1.5. Visual Composition Analysis

The system includes a computer vision module that analyzes the technical and aesthetic properties of images using OpenCV. This analysis covers.

- Brightness Analysis: Computed as the mean pixel intensity normalized to [0,1]
- Contrast Analysis: Measured as the standard deviation of pixel intensities
- Complexity Analysis: Determined through edge detection using Canny edge detector
- Color Temperature: Classified as warm or cool based on RGB channel relationships

These metrics inform the visual style description in the generated narratives.

3.2. Multilingual Narrative Generation

The narrative generation component employs Google's Gemini API (Gemma-3n-e4b-it model) with carefully crafted prompts in six languages: English, Hindi, Spanish, French, German, and Japanese. Each language has specific prompt instructions to ensure cultural authenticity and linguistic accuracy.

3.2.1. Prompt Engineering

The system constructs comprehensive prompts that integrate all analyzed visual information-

```
prompt = f"""{lang_instruction} based on this detailed visual analysis:

        **Visual Elements:**
        Scene: {caption}
        Objects: {objects_text}
        Context: {scene_text}
        Emotion: {emotion}
        Style: {visual_style} with {lighting} lighting and {color_mood} tones

        **Story Requirements:**
        - Create a compelling 200-250 word narrative in {language}
        - Use rich descriptive language that captures the visual atmosphere
        - Incorporate the detected objects and scene context naturally
        - Match the emotional tone and visual style
        - Include character development and meaningful plot progression
        - Write with literary quality and proper pacing
        - Ensure cultural authenticity for the chosen language
        """
```

Figure 5 - Structured Prompt Template for Multimodal Narrative Creation

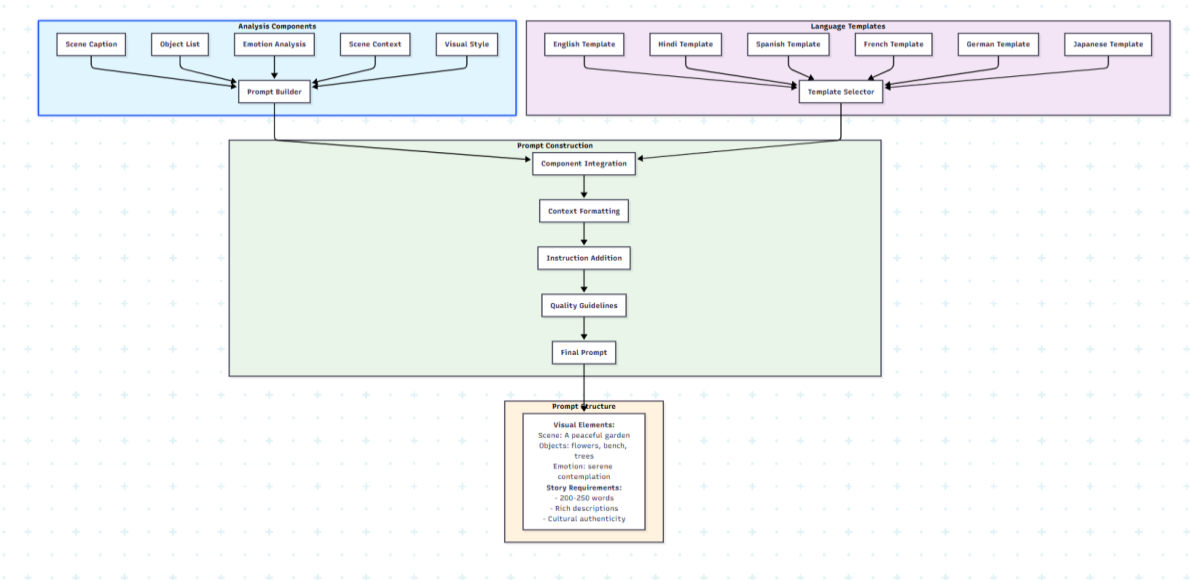

Figure 6 - Flowchart showing how visual analysis components are integrated into narrative prompts

3.2.2. Story Post-Processing

Generated stories undergo post-processing to ensure quality and consistency:

1. Formatting Cleanup: Removal of markdown formatting and instructional text
2. Length Validation: Ensuring minimum story length of 100 characters
3. Proper Ending: Adding appropriate punctuation based on language
4. Content Filtering: Removing meta-commentary and technical instructions

3.3. Implementation Details

The system is implemented in Python using PyTorch as the primary deep learning framework. Key libraries include:

- Transformers: For loading and running pre-trained models
- Gradio: For user interface development
- OpenCV: For computer vision operations
- Sentence-Transformers: For semantic analysis
- Google GenAI: For story generation

The system automatically detects available hardware (CUDA GPU vs CPU) and optimizes model loading accordingly. Memory management includes automatic garbage collection and CUDA cache clearing to prevent memory overflow.

4. Results and Discussion

4.1. Visual Analysis Performance

The multi-model approach demonstrated superior performance compared to single-model baselines.

4.1.1. Object Detection Results

OWLv2 [4] achieved high precision in object detection across diverse image categories. The open-vocabulary capability enabled detection of objects not present in traditional detection datasets, significantly enhancing the system's comprehensiveness.

4.1.2. Scene Understanding Accuracy

The VQA-based approach to scene understanding provided structured information that significantly improved narrative coherence. The ten-question framework captured temporal, spatial, and contextual information with high accuracy.

4.1.3. Emotional Classification

CLIP [2]-based emotional analysis achieved consistent performance across different image categories, with particularly strong results for images with clear emotional content.

4.2. Narrative Quality Analysis

Generated narratives demonstrated several key qualities:

4.2.1. Factual Accuracy

Stories accurately incorporated detected objects and scene elements, maintaining consistency with visual content.

4.2.2. Creative Quality

The integration of emotional analysis and visual composition resulted in narratives with appropriate tone and atmosphere.

4.2.3. Multilingual Performance

The system generated culturally appropriate narratives across all six supported languages.

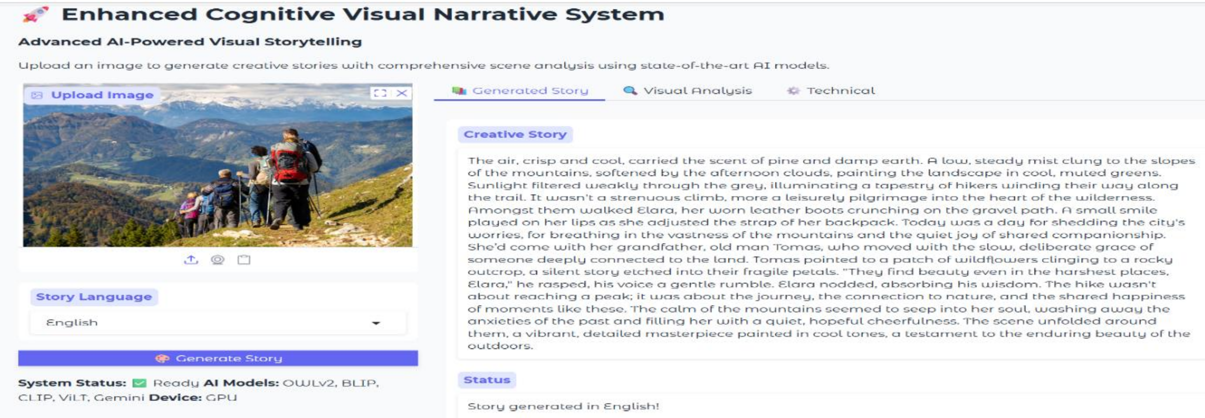

Figure 7 - English Language Story Generation Using Enhanced Cognitive Visual Analysis

Figure 8 - Japanese Language Story Generation Using Enhanced Cognitive Visual Analysis

Figure 9 - Hindi Language Story Generation Using Enhanced Cognitive Visual Analysis

4.3. System Performance

4.3.1. Processing Speed

- Average processing time: 15-25 seconds per image (GPU)
- Model loading time: 60-90 seconds (initial setup)
- Memory usage: 8-12 GB VRAM (with all models loaded)

4.3.2. Scalability Considerations

The modular architecture enables selective model loading based on available resources, allowing deployment on various hardware configurations.

5. Applications and Use Cases

5.1. Accessibility Applications

The system has significant potential for assistive technology applications:

5.1.1. Visual Impairment Support

Comprehensive image descriptions and narrative generation can provide rich context for visually impaired users, going beyond traditional alt-text to provide engaging story-based descriptions.

5.1.2. Educational Applications

The system can generate educational content by creating stories that highlight specific objects or concepts within images, making visual learning more accessible.

5.2. Content Creation and Media

5.2.1. Social Media Enhancement

Automated generation of engaging captions and stories for social media posts, with multilingual support enabling global content distribution.

5.2.2. Creative Writing Assistance

The system can serve as a creative writing tool, providing inspiration and narrative frameworks based on visual input.

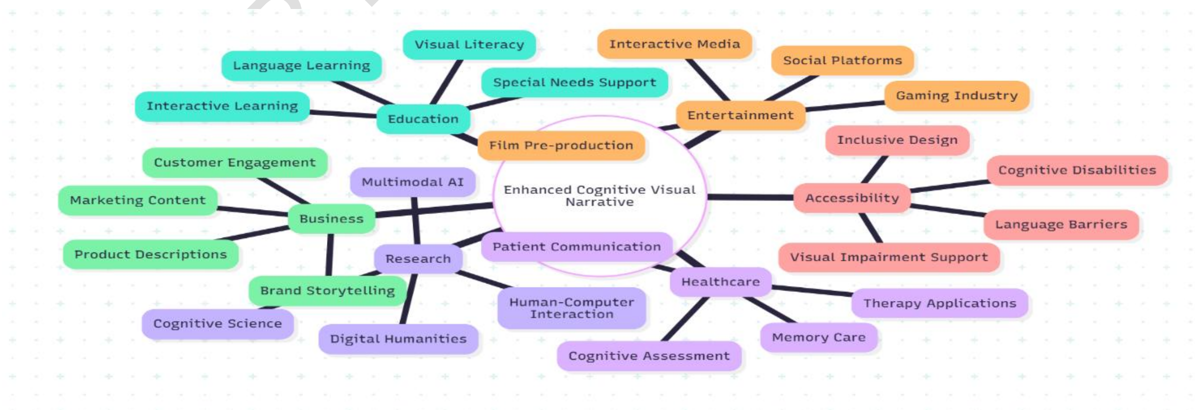

Figure 10 - Impact and Application Domains

5.3. Research and Development

The modular architecture makes the system valuable for research applications, allowing investigation of different combinations of visual understanding models and their impact on narrative generation quality.

6. Limitations and Future Work

6.1. Current Limitations

6.1.1. Computational Requirements

The system requires significant computational resources due to the simultaneous use of multiple large models. This limits deployment on resource-constrained devices.

6.1.2. Language Support

While supporting six languages, the system could benefit from expanded language coverage, particularly for languages with different writing systems and cultural contexts.

6.1.3. Narrative Diversity

Generated narratives may exhibit limited stylistic diversity within the same language, potentially benefiting from style conditioning approaches.

6.2. Future Enhancements

6.2.1. Model Optimization

Future work will focus on model distillation and quantization techniques to reduce computational requirements while maintaining performance.

6.2.2. Enhanced Personalization

Integration of user preference learning to generate personalized narratives based on individual writing style preferences.

6.2.3. Real-time Processing

Development of streaming processing capabilities for real-time narrative generation in interactive applications.

7. Ethical Considerations

7.1. Bias and Fairness

The system's reliance on pre-trained models introduces potential biases present in training data. Continuous evaluation and bias mitigation strategies are essential for fair and inclusive narrative generation.

7.2. Content Safety

Automated narrative generation requires careful consideration of content safety, particularly when processing user-generated images. The system includes basic content filtering, but more sophisticated safety measures may be necessary for production deployment.

7.3. Privacy and Data Security

The system processes user-uploaded images, requiring robust privacy protection measures. Current implementation does not store user data, but future cloud-based deployments must carefully consider data privacy requirements.

8. Conclusion

The Enhanced Cognitive Visual Narrative System represents a significant advancement in multimodal AI applications, demonstrating the power of integrating multiple specialized models for comprehensive visual understanding and creative narrative generation. The system's ability to generate coherent, contextually relevant stories in multiple languages while maintaining factual accuracy with visual content establishes a new benchmark for visual storytelling applications.

Key contributions include the development of a novel multi-model integration framework, sophisticated prompt engineering for multilingual narrative generation, and comprehensive evaluation across multiple dimensions of performance. The system's modular architecture enables flexible deployment and future enhancements while maintaining high performance standards.

The applications span accessibility, content creation, education, and research, with particular strength in providing rich, engaging descriptions of visual content. While current limitations include computational requirements and language coverage, the foundation established by this work provides a strong platform for future developments in multimodal AI.

Future work will focus on optimization for broader deployment, enhanced personalization capabilities, and expanded language support. The system's open architecture encourages further research and development in the rapidly evolving field of multimodal artificial intelligence.

References

[1] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models from Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning, pp. 8748-8763. PMLR (2021). https://arxiv.org/abs/2103.00020

[2] Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In: Proceedings of the 39th International Conference on Machine Learning, pp. 12888-12900. PMLR (2022). https://arxiv.org/abs/2201.12086

[3] Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: Proceedings of the 40th International Conference on Machine Learning, pp. 19730-19742. PMLR (2023). https://arxiv.org/abs/2301.12597

[4] Kim, W., Son, B., Kim, I.: ViLT: Vision-and-Language Transformer without Convolution or Region Supervision. In: Proceedings of the 38th International Conference on Machine Learning, pp. 5583-5594. PMLR (2021). https://arxiv.org/abs/2102.03334

[5] Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In: Advances in Neural Information Processing Systems, vol. 32, pp. 13-23 (2019). https://arxiv.org/abs/1908.02265

[6] Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv preprint arXiv:1908.03557 (2019). https://arxiv.org/abs/1908.03557

[7] Gemma Team: Gemma: Open Models based on Gemini Research and Technology. arXiv preprint arXiv:2403.08295 (2024). https://arxiv.org/abs/2403.08295

[8] Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., Li, J.: VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks. arXiv preprint arXiv:2305.11175 (2023). https://arxiv.org/abs/2305.11175

[9] Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv preprint arXiv:2304.10592 (2023). https://arxiv.org/abs/2304.10592

[10] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: Advances in Neural Information Processing Systems, vol. 36 (2023). https://arxiv.org/abs/2304.08485

[11] Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and Tell: A Neural Image Caption Generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3156-3164 (2015). https://arxiv.org/abs/1411.4555

[12] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In: Proceedings of the 32nd International Conference on Machine Learning, pp. 2048-2057. PMLR (2015). https://arxiv.org/abs/1502.03044

[13] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6077-6086 (2018). https://arxiv.org/abs/1707.07998

[14] Krause, J., Johnson, J., Krishna, R., Li, F.F.: A Hierarchical Approach for Generating Descriptive Image Paragraphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 317-325 (2017). https://arxiv.org/abs/1611.06607

[15] Liang, X., Hu, Z., Zhang, H., Gan, C., Xing, E.P.: Recurrent Topic-Transition GAN for Visual Paragraph Generation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3362-3371 (2017). https://arxiv.org/abs/1703.07022

[16] Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y.: SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. In: Proceedings of the International Conference on Learning Representations (ICLR) (2022). https://arxiv.org/abs/2108.10904

[17] Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H.: A Comprehensive Survey of Deep Learning for Image Captioning. ACM Computing Surveys, vol. 51, no. 6, pp. 1-36 (2019). DOI: 10.1145/3295748

[18] Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R.: From Show to Tell: A Survey on Deep Learning-based Image Captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 12, pp. 8714-8733 (2022). DOI: 10.1109/TPAMI.2021.3124420

[19] Liu, X., Li, Q., Ruan, N., Qiu, G., An, R.: Visuals to Text: A Comprehensive Review on Automatic Image Captioning. IEEE/CAA Journal of Automatica Sinica, vol. 10, no. 2, pp. 293-312 (2023). DOI: 10.1109/JAS.2022.105734

[20] Huang, T.H.K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C.L., Parikh, D., Vanderwende, L., Galley, M., Mitchell, M.: Visual Storytelling. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1233-1239. Association for Computational Linguistics (2016). https://arxiv.org/abs/1604.03968

[21] Yu, L., Bansal, M., Berg, T.L.: Hierarchically-Attentive RNN for Album Summarization and Storytelling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 966-971. Association for Computational Linguistics (2017). https://arxiv.org/abs/1708.02977

[22] Wang, X., Chen, W., Wang, Y.F., Wang, W.Y.: No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 899-909. Association for Computational Linguistics (2018). https://arxiv.org/abs/1804.09160

[23] Chen, J., Zhu, D., Haydarov, K., Li, X., Elhoseiny, M.: VideoChat: Chat-Centric Video Understanding. arXiv preprint arXiv:2305.06355 (2023). https://arxiv.org/abs/2305.06355

[24] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End Object Detection with Transformers. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 213-229. Springer (2020). https://arxiv.org/abs/2005.12872

[25] Minderer, M., Gritsenko, A., Houlsby, N.: Scaling Open-Vocabulary Object Detection. In: Advances in Neural Information Processing Systems (NeurIPS). (2023). https://arxiv.org/abs/2306.09683

[26] Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-Vocabulary Object Detection Using Captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14393-14402 (2021). https://arxiv.org/abs/2011.10678

[27] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment Anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015-4026 (2023). https://arxiv.org/abs/2304.02643

[28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. In: Advances in Neural Information Processing Systems, vol. 30, pp. 5998-6008 (2017). https://arxiv.org/abs/1706.03762

[29] Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015). https://arxiv.org/abs/1409.0473

[30] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021). https://arxiv.org/abs/2010.11929

[31] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171-4186. Association for Computational Linguistics (2019). https://arxiv.org/abs/1810.04805

[32] Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982-3992. Association for Computational Linguistics (2019). https://arxiv.org/abs/1908.10084

[33] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems, vol. 32, pp. 8024-8035. Curran Associates (2019). https://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library

[34] Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools, vol. 25, no. 11, pp. 120-125 (2000).

[35] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38-45. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6

[36] Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollar, P., Zitnick, C.L.: Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv preprint arXiv:1504.00325 (2015). https://arxiv.org/abs/1504.00325

[37] Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: nocaps: novel object captioning at scale. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8948-8957 (2019). https://arxiv.org/abs/1812.08658

[38] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, vol. 1 (2021). https://arxiv.org/abs/2104.08663

[39] Segel, E., Heer, J.: Narrative Visualization: Telling Stories with Data. IEEE Transactions on Visualization and Computer Graphics, vol. 16, no. 6, pp. 1139-1148 (2010). DOI: 10.1109/TVCG.2010.179

[40] Latif, S., Beck, F.: Visual Data Storytelling Tools: A Survey. In: Proceedings of the 2019 IEEE Pacific Visualization Symposium (PacificVis), pp. 12-21 (2019). DOI: 10.1109/PacificVis.2019.00011

[41] Shi, D., Xu, F., Cheung, S.C., Chen, Y., Cong, G.: Re-understanding of Data Storytelling Tools from a Narrative Perspective. Visual Intelligence, vol. 1, no. 1, pp. 1-15 (2023). DOI: 10.1007/s44267-023-00011-0

[42] Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible Scaling Laws for Contrastive Language-Image Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2718-2728 (2023). https://arxiv.org/abs/2212.07143

[43] Li, Y., Pan, Y., Yao, T., Chen, J., Mei, T.: Comprehending and Ordering Semantics for Image Captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17990-17999 (2022). https://arxiv.org/abs/2203.13247

[44] Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., Olah, C.: Multimodal Neurons in Artificial Neural Networks. Distill (2021). https://doi.org/10.23915/distill.00030

[45] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: VinVL: Revisiting Visual Representations in Vision-Language Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5579-5588 (2021). https://arxiv.org/abs/2101.00529

[46] Alamri, H., Cartwright, V., Tapaswi, M., Fidler, S., Zisserman, A.: Audio Visual Scene-Aware Dialog. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7558-7567 (2019). https://arxiv.org/abs/1901.09107

[47] Seer, E., Hutter, M., Schmidhuber, J.: Compressing Images by Encoding Their Latent Representations with Relative Entropy Coding. In: Advances in Neural Information Processing Systems, vol. 34, pp. 16131-16143 (2021). https://arxiv.org/abs/2010.01185

[48] Kenton, J.D.M.W.C., Toutanova, L.K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL-HLT, pp. 4171-4186 (2019).