International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 05 May 2026
Page No: 3835-3842
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0504

Visual-To-Text Ai Systems: Bridging Images and Content Creation

Dr M Prabu[1], Kesavan V[2], Saravanan S[3], Pravin Kumar R[4]
[1] Associate professor, Dept. of CSE, SRM Institute of Engg. & Tech., Chennai, Tamil Nadu, India
[2,3,4] UG Scholar, Dept. of CSE, SRM Institute of Engg. & Tech., Chennai, Tamil Nadu, India
EmailID: *prabum@srmist.edu.in[1], kv4787@srmist.edu.in[2], ss4309@srmist.edu.in[3], pk1079@srmist.edu.in[4]*

Abstract
With the massive growth of visual data on digital platforms, there is an increasing demand for systems capable of converting visual information into meaningful textual content. Visual to Text AI systems aim to handle the gap between images and natural language by automatically generate the description, information, and context-aware text from visual inputs. The Visual to text AI use technologies such as computer vision, deep learning, and natural language processing to understand the content and convert it into captions. This project uses Visual to text AI system combine Convolutional neural networks to extract image and transformer based model to generate text descriptions. It is used in image captioning, media systems, and for people with visual issue by describing the image. The system can create natural and easy to understand content from the image. The system helps to covert images into clear and accessible text in efficient way.

Keywords: *Visual-to-Text AI, Computer Vision, Deep Learning, Image Captioning, NLP*

1. Introduction
In today's digital world, images play a Major role in communication in areas like social media, journalism[2], and ecommerce websites to promote and express their ideas. Billions of images are created, shared, shared, and stored this produce a large number of visual data. Bit images alone do not always explain the full meaning of the content. This is the challenge for automated system to generate textual information and for people with loss of vision who rely on description to understand the content.Conversion of visual data into text has become an important area for research in artificial intelligence [3]. Visual to text AI systems are intelligent models that analyze the image and generate meaningful descriptions in natural language.Visual to text Ai systems can combine the combines computer vision and natural language processing to understand the scenes in the image and describe them in text format. Old models and methods used handcrafted features and templates to generate text which has limited scalablity and accuracy.The use of deep learning has improved massively improved both computer vision and natural language processing. Convolutional neural[1] networks have strong performance in understanding the mages and they are widely used in object detection, image classification, and image segmentation.At the same time, Recurrent neural networks and recent transformer based models like BERT and GPT have great improvement in natural language generation. The important task in Visual to text AI systems in image captioning in this the model generates short sentences that describe the content of the image [4]. Beyond simple captions, more detailed contents like stories, repots, advertisements, and educational descriptions are aimed to generate in modern systems. This progress has made it possible to use Visual to Text AI ina fields like journalism [5], content creation, marketing, and human computer interaction. Some challenges still exist in this system [6]. Systems may struggle to understand complex relationships in images, capture the exact content, produce connected texts, and avoid inaccurate descriptions. Real world images majorly have busy or messy backgrounds, different lightings, and complex details that difficult for the machine to figure out and understand. Visual to Text AI systems that combines deep visual feature extraction with the transformer based language generation models to create clear, accurate, and context aware text from the provided image. The system is designed to be scalable, adaptable to different fields, and useful for creative and assistive content creation applications

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 3835-3842
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0504

[7]. The rapid increase of digital content on social media, websites, and online ecommerce and other platforms has the need of intelligent system that can automatically understand and provide description for the image or content. Visual to Text AI system provide the solution by converting visual data into meaningful text [8]. The Visual to Text AI system combine computer vison and natural language processing to analyze the objects in the image, scenes and the relationship in the images. With the use of visual patterns related to text features, the system can generate human like content description. It turns raw visuals to understandable information making them easier for the users to access and understand. Deep learning models play an important role in the improvement of Visual to Text AI systems. Convolutional neural networks are mostly used to detect key visual feature in the provided image, like shape, color, texture, objects present in the image. These features are passed to the advanced language generation model such as transformer based architecture models, which convert the visual information into meaningful and logical text. By combing these two powerful technologies the system can effectively connect the understanding and natural language generation of the image. The amount of visual day continues to grow in everyday life, manually adding notes and descriptions to the image will be time consuming and inefficient. AI driven systems can make this process automatic and make the faster and more accurate in content generation. By joining the gap between images and textual information, Visual to Text AI system improve the way of interaction of the people with digital media. This technology also supports future developments in artificial intelligence, multimedia processing, and human computer interaction [10].

2. Literature Survey

More number of studies have explored how artificial intelligence can convert the images to textual descriptions. Early research in this field focused on basics like image annotation and tagging techniques. These methods is based on the traditional image processing and manually designed features to detect objects in the images. These approaches could recognize basics like simple visual elements, this system, often struggle to understand complex relationship between the objects and generate clear and natural language descriptions With the use and the advancement of deep learning, convolutional neural networks (CNN) have been introduced for Visual to text AI systems [9]. These models automatically learn important visual features from images without requiring manual input for the future designing. CNN-based systems can detect patterns such as texture, activities in the image, structure of the image that are present in input image to provide description [12]. This gives more accuracy for the description for the Visual to text AI systems compared to traditional image processing approaches. In addition to image recognition in converts it into text, the visual to text AI system is used for content creation and helping in digital platforms. This system use images and convert them into text and gives description using AI and natural language processing. This system helps users to automatically generate caption, simple explanation, and information from the image. It is useful for content creators, researchers, and for peoples with loss of vision to understand the content more easily. This system is mostly used in social media, and many online platforms to crate content faster [13]. This project tries to address this gap by combining the deep learning for image recognition with visual to text AI system to create content based on the image. In this system the users are allowed to upload image And the AI will analyze the image and provide description or caption for the uploaded image. This helps to create content faster without writing on their own. Improvements in visual to text Ai systems for content creation achieved by attention mechanism in image captioning models. This helps the model to focus on important parts of the image while generating the caption or description. The system can understand the scene better and generate captions. The generated content is more accurate and useful for users who want to create quick texts from for the uploaded image [11] .

3. Proposed Methodology

This section describes the methodology used for the Visual to Text AI System for Automated Content Generation shown in Figure 1.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 3835-3842
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0504

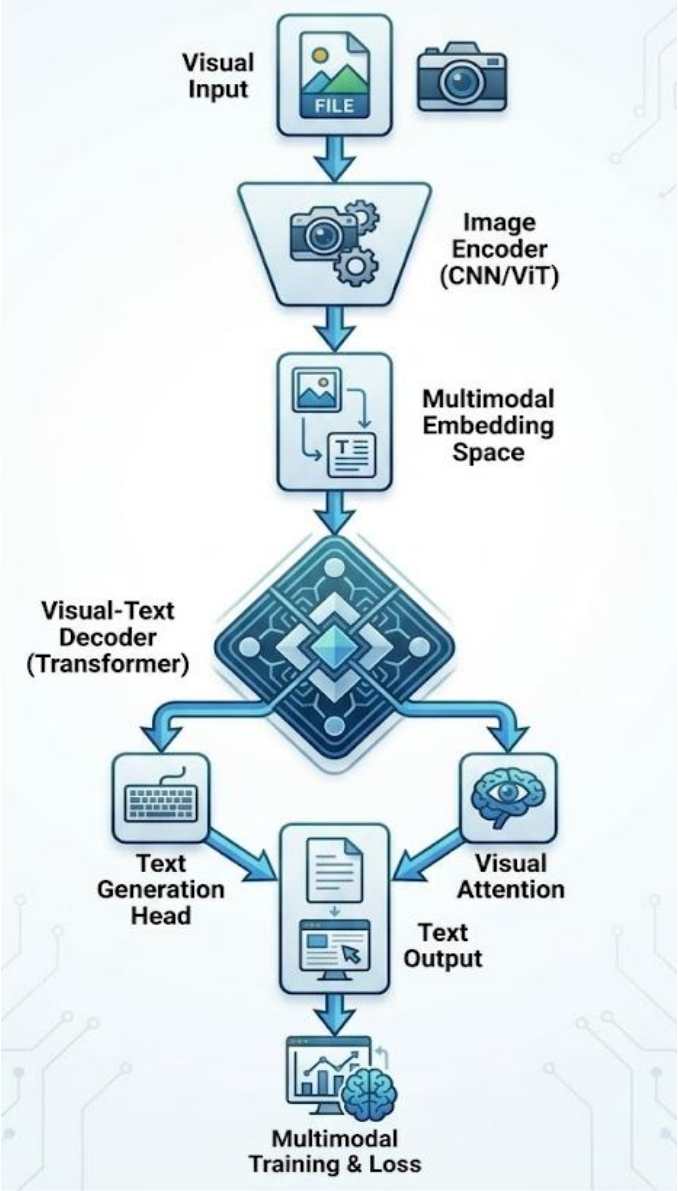

Figure 1 Proposed System Architecture of Visual to Text AI System for Automated Content Generation.

3.1.Image Input And Collection Unit

Visual to text AI system unit handles the acquisition of images from multiple sources like as digital repositories, social media datasets, image databases, and mobile camera inputs. The system supports various image formats including even JPEG, PNG, and BMP are supported by this visual to test ai model [18].

3.2.Image Enhancement and Preparation Unit

In this section, before going into the deep learning the input images given by user undergo several preprocessing operations to improve consistency and quality. The preprocessing techniques required image resizing, pixel normalization, noise removal, contrast development, and brightness adjustment for the image enhancement and preparation unit [14].

3.3.Visual Feature Extraction Engine

Tokenization converts the input text into smaller semantic units known as tokens. Tokens may represent words, subworlds, or characters depending on the tokenization strategy used by the language model. Modern LLM like byte pair encoding (BPE) or word piece tokenization. These approaches allow the system to efficiently process rare words and vocabulary [15].

3.4.Visual Understanding and Semantic Learning Unit

This unit works using a deep learning pipeline to pick important visual features from the images, It looks at different layers of the image to understand the things and objects, shapes, colors, and textures present in the images to create suitable text captions and descriptions. This model gradually learns how to convert image into meaningful description for the users [16].

3.5.Deep Learning Model Design

Different deep learning architectures were considered and evaluated to achieve accurate visual to text conversion. The model design concentrates on productive visual feature extraction and natural language generation. Various CNN and transformer based on the architectures were studied based on performance effectiveness, computational complexity, and ability of multimethod learning. The most suitable architecture was selected to assure accurate visual interpretation and fluent text generation for deep learning model design [17].

3.5.1.CNN-Based Feature Extraction Model

A filter based neural network architecture was designed to obtain significant visual features from images. The model consists of different convolution layers followed by collecting layers and fully connected compact layers. Fixed linear unit activation functions are used to introduce irregular pattern and improve learning capability. The starting layers of the convolutional neural network took low level features such as edges, colors, and textures.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 3835-3842
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0504

medium layers identify object parts and shapes, while deeper layers identify complex visual patterns and scene structures. These ranked features provide a detailed representation of the image, which is used by the language generation model to produce a detailed text. The final feature list extracted by the CNN is passed to the text generation module. Merging layers are used to reduce the physical dimensions of the feature maps, and decreasing algorithmic complexity and blocking overfitting. Max pooling will help to keep the most important feature from image and reduce the unnecessary information [19].

3.5.2. Transformer-Based Language Generation Model

To convert the visual features into meaningful text, this system use the transformer based language model this will create meaningful natural language description of the images. Transformers have become the leading architecture in modern Natural Language Processing (NLP) tasks due to their capability to capture situational relationships between words using concentration mechanisms[23]. In the Visual to text AI system proposed framework, the visual feature vector generated by the CNN coder is provided as input to the transformer model. The transformer consists of multiple encoder and decoder layers that process information using self attention and feed forward neural networks. The self attention mechanism allows the model to understand relationships between different words in a sentence and to maintain context based clarity throughout the generated text [24]. The attention mechanism also enables the system to focus on the most applicable parts of the image during text generation. For example, if the image contains multiple objects, the model flexibly assigns attention weights to different visual features while generating each word in the caption [22].

3.5.3. Content Expansion and Enhancement Model

Few-shot prompt engineering is the main central system of the proposed system. Instead of building a new model and training that model completely and separately for each particular task for classification [25], this system leverages and gives a pre-trained large language model and tell the LLM what to do by giving prompts. This prompt which are given to the LLM contains examples which are labeled that demonstrate the relationship between input and text classification pattern without requiring huge training and data. This deep architecture demonstrates strong performance in identifying visually complex temple structures and distinguishing similar architectural styles across different dynastic influences [20].

3.6.Model Training and Validation

The model used in the visual to text AI systems is trained using a large collection of images paired with text descriptions. the datasets are divided separately into training, validation, and testing sets. The training set in model training is used to update the model weights and improve leaning, while the validation set is used in monitoring performance and prevent the overfitting. Data improvement techniques such as image rotation, scaling, flipping, cropping, and brightness variation were applied to improve the overview ability of the model. These techniques help the model learn durable visual features and perform well on varied real world images. The training process used categorical information loss to evaluate prediction errors. Optimization algorithms such as Adam optimizer were used to update the model variables. Early stopping techniques were applied to stop training once the validation performance stabilized, preventing overfitting and ensuring better generalization [21].

3.7.Model Evaluation and Performance Metrics

To ensure accurate evaluation of the visual to text generation system, different performance metrics were used to measure the quality of generated text in the system [26].

- Classification Accuracy: The Accuracy is a factor that determines the correctness of the system in creating captions according to the image that contains the reference descriptions of the listed dataset [27].
- Precision: Precision measures the number of the generated words or the produced words that are relevant and correct in comparison with the real image content that is given by user.
- Recall: Recall is used to determine stability and capacity of the model to incorporate all the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 3835-3842
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0504

significant text and details found within the image uploaded in the generated text.

- F1-Score: The F1 score combines precision and recall into a single evaluation metric, providing a balanced assessment of the caption generation performance.
- Confusion Matrix Analysis: The similarity of generated captions and descriptions made by human beings was checked by semantic similarity measures such as BLEU and METEOR scores. Such measures are used to identify the extent to which the text is similar to the descriptions of the ground truth.
- Inference Time and Model Efficiency: To ascertain the efficiency of the system in the application of real-time, the inference time of the model was checked. Propagation Approved models will permit quicker capturing and they may be applied in systems such as mobile applications, assistive technologies, and web media. The suggested approach of Visual to Text AI System will integrate image-preparation, visual- features propagation in the form of deep learning, and language-generation by the use of transformers will transform images into meaningful textual messages. It permits efficient and successful generation of the contents and increases accessibility to the visually impaired consumers and facilitates multimedia analysis intelligently.

4. Results and Discussion

The proposed Visual to Text AI system was developed and tested using a dataset contains images along with their descriptive captions. The system combines convolutional to extract important neural networks for visual feature extraction and transformer-based models for text generation. The main goal of this evaluation was to examine how effectively the system can produce accurate and make sense in the content of the text based visual inputs [29].

4.1.Model Performance

The Image datasets were used as the training procedure in which there are diversity of objects, scenes and activities. In order to increase the generalization ability and the strength of the model, there were various data augmentation methods that were used in the training process. Such methods were rotation, horizontal flipping, scaling, and brightness. Data augmentation assisted the model to learn under various image conditions and cut off the chances of overfitting. The CNN encoder was useful in the training and validation phases to extract hierarchical visual features of the input images. The initial layers of the convolution were used to detect simple features of the objects like edges, texture, and color pattern, whereas the deeper layers learned intricate features like the type of object and the general view of the scene. These features that were extracted were later fed to the transformer based language generation module. The transformer architecture demonstrated stable learning throughout the training and acquired accuracy in the generated topics slowly. The attention system assisted the model in paying attention to similar and relevant areas of the image when each word of the sentence was generated. This significantly enhanced the quality of the context generated captions. The loss analysis presented above indicated the ability to converge during training and indicated that the model parameters were optimized effectively. The validation accuracy was maintained during the training process and this shows that the model had the capability of generalizing well on images that were not in the training [28].

4.2.Output Prediction

The output prediction stage will look at the extent to which the system is able to convert the visual inputs to meaningful textual descriptions. Upon uploading of an image by a user, the preprocessing module first of all standardizes the image format and then resizes the image and applies noise removal procedures to it. The resultant processed image is given to the CNN based visual encoder to extract features. Transformer based language generation model takes the obtained visual feature vectors as input. The transformers generate the most logical sequence in the analysis of the visual features and the previous already generated words. The emphasis on mechanism enables the model to dynamically attend to various components of the picture as it produces every word of the sentence [30]. The end product that the system produces consists of the descriptive captions that

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 3835-3842
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0504

summarize the visual contents of the image. As an example, images with human activities can create captions such as A group of people playing football on a grassy field, whilst object focused on images can have such captions as A red car parked near a building. These findings demonstrate the system capability of analyzing visual scenes and compose them in natural language. The generated captions can equally be converted to textual description material in detailed content with the assistance of the content enhancement module. The capability places the system into a position where it is capable of making detailed textual descriptions that can be used in applications such as automated content, social media captioning, digital journalism, and multimedia documentation. It has been experimentally observed that the system is effective in cases of familiar objects and structured scenes recognition. There were minor errors where there were complicated backgrounds of images, objects overlapping, and abstract images. The more the training data, the better its accuracy in prediction will be as more features will be used to represent it shown in Figure 2-4 .

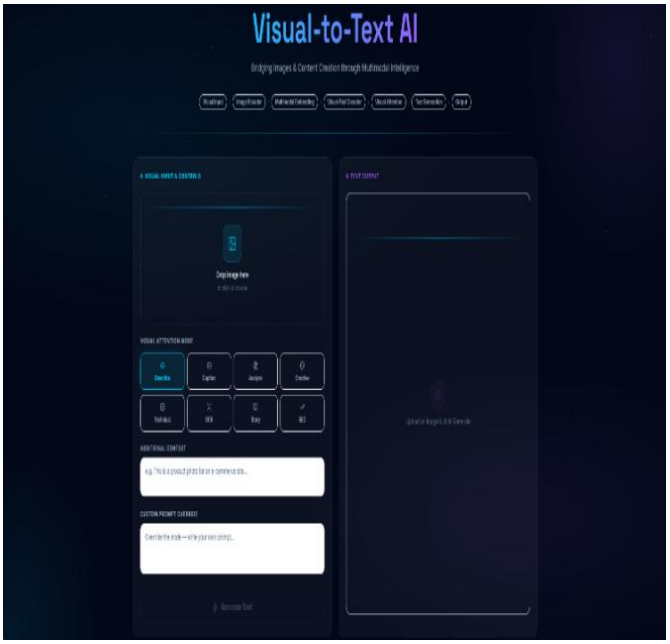

Figure 2 User Interface of the Visual to Text AI system

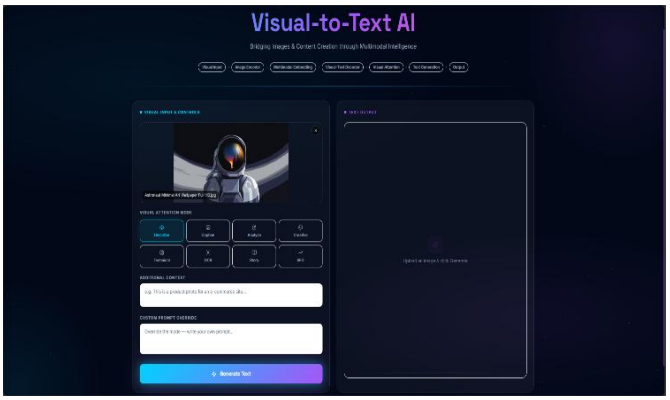

Figure 3 Image Uploaded As Input For Text Generation

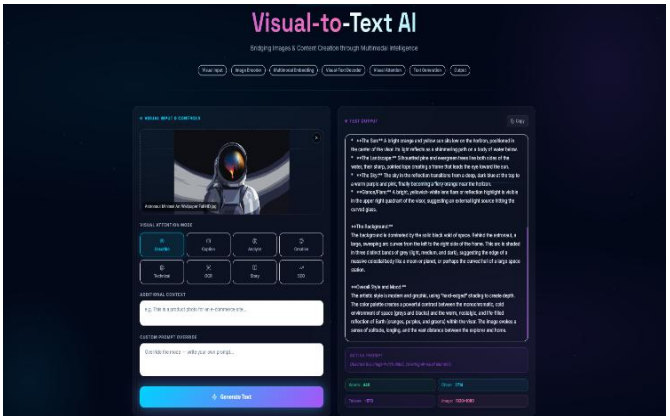

Figure 4 Generated Text Output for the Input Image

Conclusion

The Suggested Visual To Text Ai System Is Effective In Providing The Visual Data With Textual Content Generation Due To Combining Both Computer Vision And Natural Language Processing. It Has Been Made Possible To Produce Accurate And Informative Captions Which Represent The Visual Content Of The Images Accurately And At The Same Time Guarantees Consistent Model Performance As Determined By The Results Of Experiments. The Approach That Significantly Reduces The Need For A Manual Image Annotation And Supports Automated Content Generation In Various Domains Like Digital Media Management, Tools For Multimedia Documentation, And Intelligent Human-Computer Interaction. Overall, The Proposed Framework Demonstrates The Potential Of Multimodal Ai Systems To Convert Visual Data Into Meaningful Textual Information

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 3835-3842
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0504

And Offering A Scalable Approach For Ai-Driven Content Creation Applications.

Acknowledgment

We would like to express our gratitude to our project guide, Dr. M. Prabu, for his constant support, helpful guidance, and direction throughout the duration of this project. His guidance kept us on track and enabled us to finish our work successfully. We also thank the faculty and fellow students at SRM IST, Ramapuram, for their useful suggestions and support whenever required. A special acknowledgment goes to our families and friends for their support and encouragement. We would finally appreciate the open-source community for supplying the tools and resources used so that we are able to make this research work.

References

[1]. S. Mystakidis, "Metaverse," Encyclopedia, vol. 2, no. 1, pp. 486–497, 2022.

[2]. S. Sai, D. Goyal, V. Chamola, and B. Sikdar,"Consumer electronics technologies for enabling animmersive metaverse experience," IEEE Consum.Electron. Mag., vol. 13, no. 3, pp. 16–24, May 2024.

[3]. S. Sai, A. Garg, and V. Chamola, "Navigating themetaverse: A comprehensive analysis of consumerelectronics prospects and challenges," ACM Trans.Internet Technol., 2024.

[4]. I. L. Chamusca, C. V. Ferreira, T. B. Murari, A. L.Apolinario Jr, and I. Winkler, "Towards sustainable virtual reality: Gathering design guidelines for intuitiveauthoring tools," Sustainability, vol. 15, no. 4, 2023,Art. no. 2924.

[5]. V. Krau, A. Boden, L. Oppermann, and R. Reiners,"Current practices, challenges, and designimplications for collaborative AR/VR applicationdevelopment," in Proc. 2021 CHI Conf. Hum. Factors Comput. Syst., 2021, pp. 1–15.

[6]. J. Ratican, J. Hutson, and A. Wright, "A proposedmeta-reality immersive development pipeline:Generative AI models and extended reality (XR)content for the metaverse," J. Intell. Learn. Syst. Appl.,vol. 15, pp. 24–35, 2023.

[7]. S. Palani and G. Ramos, "Evolving roles and workflowsof creative practitioners in the age of generative AI," inProc. 16th Conf. Creativity Cogn., 2024, pp. 170–184.

[8]. A. H. Hwang, "Too late to be creative? ai-empoweredtools in creative processes," in Proc. 2022 CHI Conf.Hum. Factors Comput. Syst. Extended Abstr., 2022,pp. 1–9.

[9]. Y. Tewel, Y. Shalev, I. Schwartz, and L. Wolf,"ZeroCap: Zero-shot image-to-text generation forvisual-semantic arithmetic," in Proc. IEEE/CVFConf. Comput. Vis. Pattern Recognit., 2022,pp. 17918–17928.

[10]. W. Zhang et al., "Relational graph learning forgrounded video description generation," in Proc.28th ACM Int. Conf. Multimedia, 2020,pp. 3807–3828.

[11]. T. Tahara, T. Seno, G. Narita, and T. Ishikawa,"Retargetable AR: Context-aware augmented reality inindoor scenes based on 3D scene graph," in Proc.2020 IEEE Int. Symp. Mixed Augmented RealityAdjunct, 2020, pp. 249–255.

[12]. S.Buongiorno and C. Clark, "Leveraging gaming toenhance knowledge graphs for explainable generativeAI applications," in Proc. 2024 IEEE Conf. Games,2024, pp. 1–4..

[13]. N. Shabani et al., "Attention-based graphsummarization for large-scale information retrieval,"IEEE Trans. Consum. Electron., early access, Jun. 11,2024,

[14]. P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learningwith transformers: A survey," IEEE Trans. Pattern Anal.Mach. Intell., vol. 45, no. 10, pp. 12113–12132, Oct.2023.

[15]. V. Beltran, J. C. Caicedo, N. Journet, M. Coustaty, N. Journet, M. Coustaty, F. Lecellier, and A. Doucet, "Deep multimodal learningfor

[16]. Ji et al., "CRET: Cross-modal retrieval transformerfor efficient text-video retrieval," in Proc. 45th int. ACMSIGIR Conf. Res. Develop. Inf. Retrieval, 2022,pp. 949–959.

[17]. W. Zhang and Y. Wang, "An empirical study of theimpact of metaverse storytelling on intentions to visit,"Inf. Technol. Tourism, vol. 25, no. 3, pp. 411–432,2023.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 3835-3842
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0504

[18]. A. I. Alhussain and A. M. Azmi, "Automatic storygeneration: A survey of approaches," ACM Comput.Surv., vol. 54, no. 5, pp. 1–38, 2021.
[19]. A. Yuan, A. Coenen, E. Reif, and D. Ippolito,"Wordcraft: Story writing with large language models,"in Proc. 27th Conf. Intell. User Interfaces, 2022,pp. 841–852..
[20]. J. Freiknecht and W. Effelsberg, "Proceduralgeneration of interactive stories using language
[21]. C. Saharia et al., "Photorealistic text-to-image diffusionmodels with deep language understanding," Adv.Neural Inf. Process. Syst., vol. 35, pp. 36479–36494,2022.
[22]. N. Singhal, P. P. Singh, N. Singh, M. Singh, and H.Singh, "Text to video using GANs and diffusionmodels," Jordanian J. Comput. Inf. Technol., vol. 10,no. 2, pp. 91–106, 2024.
[23]. A. Raj et al., "Dreambooth3D: Subject-driven text-to-3D generation," in Proc. IEEE/CVF Conf. Comput. Vis.,2023, pp. 2349–2359.
[24]. C.-H. Lin et al., "Magic3D: High-resolution text-to-3Dcontent creation," in Proc. IEEE/CVF Conf. Comput.Vis. Pattern Recognit., 2023, pp. 300–309.
[25]. Z. Yin, Y. Wang, T. Papatheodorou, and P. Hui,"Text2VRScene: Exploring the framework ofautomated text-driven generation system for VRexperience," in Proc. IEEE Conf. Virtual Reality 3DUser Interfaces, 2024, pp. 701–711.
[26]. I. K. Raharjana, D. Siahaan, and C. Fatichah, "Userstories and natural language processing: A systematicliterature review," IEEE Access, vol. 9, pp. 53811–53826, 2021.
[27]. A. Almarzouqi, A. Aburayya, and S. A. Salloum,"Prediction of user's intention to use metaverse systemin medical education: A hybrid SEM-ML learningapproach," IEEE Access, vol. 10, pp. 43421–43434, 2022.
[28]. V. Chamola et al., "Beyond reality: The pivotal role ofgenerative AI in the metaverse," IEEE Internet ThingsMag., vol. 7, no. 4, pp. 126–135, Jul. 2024.
[29]. V. Chamola et al., "Beyond reality: The pivotal role ofgenerative AI in the metaverse," IEEE Internet ThingsMag., vol. 7, no. 4, pp. 126–135, Jul. 2024.
[30]. M. Iman, H. R. Arabnia, and K. Rasheed, "A review ofdeep transfer learning and recent advancements,"Technologies, vol. 11, no. 2, 2023, Art. no. 40.