iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78948

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Intelligent Multi-Model Visualisation Framework Using Large Language Models

Rupsa Das[1], Nainsi Jaiswal[2], Tejasvini Ghadage[3], B. Pooja[4], Trupti Cariappa[5], Prof. Debasmita Datta[6]
Dept. of Computer Application, Sri Balaji University, Pune, Maharashtra, India

Abstract: ***The growth and spread of huge, non-homogeneous datasets in areas like healthcare, finance, and education has also created significant difficulties in providing data in a way that is useful enough to be accessible to non-expert stakeholders. Traditional methods of visualization, such as fixed dashboard displays and ready-made charts, have been sufficient as a communication mechanism to trained analysts but often fail when communicating subtle information to users not prepared in the formal data science. This paper introduces Intelligent Multi-Modal Visualization Framework using Large Language Models (IMVF-LLM), an end-to-end model that combines multimodal data integration, chain-of-thought style reasoning that relies on Large Language Models and automated declarative visualization generation to convert heterogeneous data sets into coherent human-readable narrative reports. The framework relies on the state-of-the-art vision-language models including GPT-4, CLIP, Flamingo, BLIP-2, and MiniGPT-4 to do alignment between the vision and the visualization synthesis. An organizational user study of 214 subjects, when controlled showed an increase by 71 percent in data understanding and an 18 percent enhancement in information storage as compared to either a visual or textual display. Experimental assessment also provided 22 per cent enhancement in the cross-modal retrieval accuracy and 92 per cent fidelity in generated Vega-Lite specifications over human generated counterparts. Taken together, these results justify the enhanced generalizability of multimodal storytelling that is based on LLM to analyze data in a variety of real-life applications.***

Keywords: Multimodal AI; Large Language Models; Data Storytelling; Data Visualization; Human-Computer Interaction; Text-to-Speech Synthesis; Cross-Modal Learning.

I. INTRODUCTION

In the digital age, there is an unprecedent volume of large-scale data that is produced in the healthcare, finance and education industries. Despite the significant value of such data in the context of organizations and people, it cannot be used with much utility in the absence of special tools that would facilitate its analysis and interpretation. Conventional visualization methods, such as static dashboards and pre-computed charts, prove useful in revealing analytical information, but their results often require technical skills, and, therefore, cannot be used by the non-technical population. This chronic discrepancy highlights the importance of systems that can make big data interpretable in an intuitive and adjustable form. However, recent developments in multimodal machine learning and Large Language Models (LLMs) have provided new opportunities in terms of combining various sources of information, including text, images, audio, and structured tabular data. Use of vision-language models including CLIP [1], Flamingo [2], BLIP-2 [3] and MiniGPT-4 [4] have proven to have a significant ability to bridge the gap between the visual and textual space, with complex cross-modal reasoning being possible. At the same time, the current literature on the topic of narrative visualization and data storytelling [6] supports the substantial evidence that contextual narrative can be stimulated by using structured visual representations and lead to a significant increase in user engagement and data understanding.

Based on these complementary developments, this article presents the Intelligent Multi-Modal Visualization Framework using Large Language Models (IMVF-LLM): an end-to-end system that combines multimodal data fusion, LLM-based reasoning, and automated visualization generation to generate comprehensive storytelling experiences. The architecture is aimed at democratizing data analytics by producing visual, written and audio stories that enable users of all levels of technical skill to interact productively with complicated data.

II. RELATED WORK

A. Multimodal Machine Learning Foundations Multimodal

Machine learning is a relatively new field that began to emerge in 2007, with key research contributions emerging from 2008 through 2013.<|human|>A. Foundations of Multimodal Machine Learning Multimodal machine learning is a relatively recent subfield that started to gain traction starting in 2007, however, with main research contributions being made between 2008 and 2013.

©IJRASET: All Rights are Reserved

The first major advancement was in multimodal machine learning where the main problem problems were represented, aligned, and translated as well as fused and co-learned heterogeneous streams of data. Baltrušaitis et al. [7] have offered a detailed taxonomy to describe the process and relation of modalities including vision and language and the weakness of unimodal methods in noisy conditions. Ngiam et al. [8], proposed multimodal deep learning schemes that learn as a unified learning approach in response to audio-visual inputs and proved to be more robust with common representations. This was furthered by Srivastava and Salakhutdinov [9] with deep Boltzmann machines, which can fuse probabilistically unsupervised, with no labeled data. Although these foundations had developed both early and late fusion strategies, they also demonstrated the necessity of intermediate strategies that could represent more intricate cross-modal dependencies.

B. Developments in Data Fusion and Cross-Modal Learning.

The concept of multimodal data fusion has developed significantly as a complex feature concatenation to intricate cross-modal systems. Chandrasekaran et al. [14] investigated cross-modal learning in semantic analysis, making it possible to transfer knowledge to different modalities to improve the interpretation. To reveal latent patterns, Zhang et al. [13] presented interactive visual analysis techniques, which combine the multimodal fusion in the analysis with the user-directed exploration. The cross-modal alignment proved to be an important ability, where models trained on joint embedding spaces to overcome the representational gap between vision and language, and thus the paradigms of vision-language that textual supervision induces visual interpretation.

C. Vision-Language Models

Vision-language models are a significant step in multimodal integration. The CLIP model by Radford et al. [1] positioned images and text in a common representational space through contrastive learning, which allowed it to perform zero-shot in a variety of downstream tasks.

It is based on this that the Flamingo introduced by Alayrac et al. [2] proposed few-shot visual language modelling with non-autoregressive language models that pre-process visual features. The BLIP-2 [3] bootstrapped language-image pretraining of Li et al. [3], learns to effectively bridge the modality gap with a small querying transformer and achieves significantly better results on vision-language benchmarks than large models.

A MiniGPT-4 designed by Zhu et al. [4] was able to align frozen visual encoders with state-of-the-art LLMs through a single projection layer. These capabilities were extended to native multimodal processing capabilities by OpenAI with GPT-4 [5], which shown to be able to reason intelligently on combined text and image inputs.

D. Data Visualization, Narrative and Storytelling.

Similar advances in the study of visualization focus on narrative form and the design of people. Segel and Heer [6] described narrative visualization in author-oriented and reader-oriented continuums, and found genres such as annotated charts and interactive timelines. Knaflic [11] emphasized the conversion of unprocessed numerical data into convincing stories with a regular scale and framing that reduce the cognitive load. Hullman and Diakopoulos [10] have examined the sphere of visualization rhetoric and have shown that the design decisions determine the interpretation in the similar manner to persuasive communication. Surveying visualization related to human-centered visualization, Wang et al. [12] applied machine learning to map and produce insights without neglecting the needs of users (constantly). The works taken as a whole emphasize the significance of effective communication but pay little attention to modalities alone.

E. Cons of Existing Systems and Motivation.

Even though there is a serious improvement, the current multimodal systems have serious weaknesses in intelligent and automatic processing of various kinds of data to be visualized. Recent multimodal LLMs have difficulties with structured semantic encodings like data-to-map chart mappings, restricting their ability to extract insights and rebuild visual representations. Semantic understanding, interpretability, cross-modal integration, and underrepresented data distributions continue to have serious gaps. Conventional visualization systems also demand a lot of manual effort in the context of building the narrative, with no automatic logic to perform insight generation.

All of these restrictions inspire the creation of an LLM-based model that would combine high-level multimodal pretraining with visualization generation to make the content intelligible and comprehensible through analysis.

.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

III. PROBLEM STATEMENT

The existing systems of multimodal visualization have serious drawbacks in their ability to close the gap between the interpretation of data and its efficient visualization. Although vision-language models like CLIP [1] and GPT-4 [5] are high-performance versions of interpreting visual input, they do not decode structured encodings in charts and often do not produce interpretable narratives out of fused representations, leading to superficial analytical insights.

There is a very serious gap between multimodal understanding and intelligent visualization: current systems combine heterogeneous data but do not contain the reasoning provided that can automate the process of making insights or dynamically change the visual presentation. Among them are: (1) visual pattern recognition that cannot be processed by the textual reasoning module; (2) semantic mismatch that introduces representational biases; (3) low interpretability of intermediate fusion processes; (4) the lack of automated generation pipelines, thus necessitating the use of expensive manual ones. The IMVF-LLM framework resolves these gaps by using chain-of-thought reasoning, tool orchestration and using natural language interfaces to allow end-to-end interpretation, reasoning and visualization across heterogeneous multimodal data.

Table I. Capability comparison of existing systems vs. IMVF-LLM

System	*Align.*	*CoT*	*Auto viz*	*NL int.*	*Struct. data*
CLIP [1]	✓	✗	✗	✗	✗
Flamingo [2]	✓	*Part.*	✗	*Part.*	✗
BLIP-2 [3]	✓	*Part.*	✗	*Part.*	✗
MiniGPT-4 [4]	✓	*Part.*	✗	✓	✗
GPT-4 [5]	✓	✓	*Part.*	✓	*Part.*
IMVF-LLM (ours)	✓	✓	✓	✓	✓

✓ = full ✗ = none Part. = partial Blue row = proposed system

IV. PROPOSED METHODOLOGY

The suggested IMVF-LLM combines multimodal fusion with the reasoning of LLM-based automated and interpretable visualization generation. The structure consists of six functional layers as follows.

A. Data Input Layer

The input layer is used to process heterogeneous data modalities, which include: (1) unstructured text and natural language descriptions and user queries; (2) images and pre-existing charts; (3) audio in the form of transcripts that are processed through preprocessing pipelines; and (4) structured data, which can be in the form of CSV and JSON. Normalization of all modalities is done by using modality specific preprocessing to allow common downstream processing.

B. Multimodal Feature Extraction Layer.

Embeddings, which are modality-specific, are obtained with pretrained encoders: a frozen ViT (backend of BLIP-2) [3] with visual encoders, BERT-style encoders with textual encoders and spectrogram-based encoders with audio. The layer is based on effective querying mechanisms of BLIP-2, which minimizes the computational burden whilst maintaining cross-modal discriminating characteristics.

C. Cross-Modal Alignment Module.

Each of the individual modalities is encoded to feature representations which are in a common embedding space aligned by contrastive learning goals based on CLIP [1] and Flamingo [2]. A semantic information transfer across data type and cross-modal generalized inference is made possible through a lightweight projection layer similar to the one used in MiniGPT-4 [4], which fills the modality gap.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

D. LLM Reasoning Engine

The logic processor is the backbone of the core reasoning which uses a GPT-4-based backbone [5] to read fused multimodal representations. Chain-of-thought (CoT) prompting directs the semantic analysis, pattern recognition and insight extraction. This engine specifically focuses on this question by overcoming the visualization decoding deficiencies in previous research through reasoning on structured encodings and producing interpretable texts to summarize visual data.

E. Visualization Generation Module.

The LLM reasoning engine runs a downstream visualization pipeline which generates declarative chart specifications in Vega-Lite grammar, which is run through the Altair Python library. Narrative forms are defined by Segel-Heer visualization genres [6] that allow reader-driven and author-driven modes of storytelling.

F. User Interaction Layer

Natural language queries, iterative feedback loops and rhetorical modifications to generated visualizations are supported by the interaction layer. This element realizes the principles of human-centered design [12] by allowing users to optimize outputs with the help of conversational dialogue, thus addressing data literacy levels of various levels.

G. System Workflow

The processing pipeline works in the following way: (1) the input layer takes the end-to-end heterogeneous multimodal data and preprocesses it; (2) features representations are extracted and normalized with the help of modality-specific encoders; (3) the cross-modal alignment project maps embeddings to a common latent space; (4) the LLM reasoning engine semantically analyzes and produces visualization specifications; (5) Altair renders declarative chart code; and (6) the user can interact with the generated results by using natural language queries, which will trigger Under the experimental hardware setup of Section V, end-to-end query latency was observed to be 15 seconds on average.

V. IMPLEMENTATION

The IMVF-LLM was introduced as a pipeline based on the multimodal machine learning with the visualization generation using LLM. The implementation was done on PyTorch 2.0 using HuggingFace Transformers and tested on a cluster of NVIDIA A100 GPUs. The heterogeneous inputs are preprocessed by data ingestion with modality-specific encoders: images and videos are encoded through a frozen Vision Transformer (ViT) which is part of BLIP-2 [3]; text and structured information is encoded using the tokenizer of GPT-4 [5]; audio data is extracted with spectrograms compatible with audio adaptations of CLIP [1].

Multimodal fusion makes use of cross-modal attention based on the original research of Baltrusaitis et al. [7], which projects embeddings into a common latent space using a lightweight query transformer simulating MiniGPT-4 [4]. The present design assists in transferring knowledge cross-modally, as developed on the methodology of Ngiam et al. [8] and Srivastava and Salakhutdinov [9]. These fused representations are then introduced to a central GPT-4 backbone that is encouraged to identify patterns, suggest correlations, and write narratives summary. The visualization pipeline communicates with the Vega-Lite declarative grammar, and is coordinated by the LLM reasoning engine. Natural language user queries (e.g., Visualize climate trends from satellite images and reports) are used to generate Vega-Lite specifications, which are rendered with Altair in Python using the LLM. The interactive storytelling is based on the narrative visualization model of Segel and Heer [6] with the application of human-centered design concepts of Wang et al. [12]. The datasets assessment was carried out on the basis of synthetic multimodal benchmarks (Visual Genome with added tabular metadata) and real-world datasets such as COCO with added text annotations [13][14].

Table II. Implementation configuration summary

Component	Specification
GPU	NVIDIA A100 (80 GB), cluster
Framework	PyTorch 2.0, HuggingFace Transformers
Vision encoder	Frozen ViT — BLIP-2 [3]
LLM backbone	GPT-4 (OpenAI API) [5]

Component	Specification
Audio encoder	Spectrogram, CLIP-aligned [1]
Viz library	Altair / Vega-Lite (Python)
Fusion	Cross-modal attn. + query transformer [4]
Benchmarks	Visual Genome, COCO [13][14]
Query latency	~15 s (end-to-end)

attn. = attention. Hugging Face = Hugging Face Transformers library.

VI. RESULTS AND DISCUSSION

IMVF-LLM has good performance on various evaluation dimensions. The framework outperformed standalone baselines - CLIP [1] and Flamingo [2] by 22 percent (mAP@5: 0.78 vs. 0.64). The system demonstrated an accuracy of 85% in detecting anomalies using LLM-based reasoning on a multimodal benchmark of Visual Genome images with synthetic tabular data, more than unimodal LLMs, by relying on bootstrapped pretraining of BLIP-2 [3]. Incorporation of miniGPT-4 [4] extended instruction-following visual instruction to 92% fidelity in generated Vega-Lite specifications compared to specifications written by humans.

This was improved dramatically by the use of narrative overlays that describe generated learnings in natural language (e.g. Satellite imagery correlates 0.87 with textual drought reports), and in keeping with the principles of storytelling espoused by Knaflic [11]. IMVF-LLM visualizations were rated as clear by N=25 data scientists with a 4.7/5 rating, as opposed to 3.2/5 rating of manually designed D3.js equivalents, which confirms the storytelling effectiveness of the framework [6]. Cross-modal fusion also made possible new insights to be made, like audio-visual sentiment alignment (F1: 0.82), the one that would not be achieved in unimodal systems.

An independent controlled user study of 214 participants showed a 71% better understanding of the data and an 18% better storage than either text or visual representation, and significantly better engagement among all groups of users. The multimodal storytelling evaluation metrics Wang et al. [12] developed with a human focus established a 30-percent decrease in time taken to extract the insight, which highlights the cognitive advantages of the specified approach.

Result discussion also presents these results in the context of the multimodal literature of learning: GPT-4 reasoning offers [5] enhancements to the zero-shot representational capabilities of CLIP, allowing it to generate stories on unstructured data. All these have been the results of the potential of IMVF-LLM as an interpretable scalable analysis platform that bridges persistent.
interactive multimodal system gaps known by Baltrusaitis et al. [7].

Table III. Quantitative performance — IMVF-LLM vs. baselines

Metric	IMVF-LLM	Baseline	Improvement
mAP@5 retrieval	0.78	0.64	+22%
Anomaly detection	85%	63%	+22pp
Vega-Lite fidelity	92%	—	vs. human
User clarity (/5)	4.7	3.2	+47%
Comprehension gain	+71%	—	N=214
Retention gain	+18%	—	N=214
Insight time	−30%	—	faster
Sentiment F1	0.82	N/A	unimodal—absent

Baseline = best of CLIP/Flamingo. — = not applicable. pp = percentage points.

VII. LIMITATIONS

Although IMVF-LLM has proven to be an excellent empirical performer, it has a number of significant limitations that limit its applicability in practice. Scalability is one of the key issues: work with datasets of 10,000 and more image-text pairs can consume up to 48 GB of VRAM on a single GPU, which does not allow running it in real-time as compared to smaller models like CLIP [1]. Iterative LLM inference adds new computational expenses, and has average scale costs per query of about 0.05, which is prohibitive to edge computing applications.

Web-scraped pretraining corpora have inherited biases on system performance with underrepresented data types. The performance of BLIP-2 [3] and MiniGPT-4 [4] on non-English audio inputs is worse in terms of accuracy (reduction of 15%), and the Flamingo-style few-shot learning [2] reduces but does not eliminate cultural biases when generating visualization stories. Interpretability is partially true: although the chain-of-thought explanations improve user trust, cross-modal fusion layers with black boxes obscurantism reflects the fears of Baltrušaitis et al. [7]. Lastly, narrative generation may be oversimplified in case of data uncertainty, as found by Hullman and Diakopoulos [10] and Segel and Heer [6].

VIII. FUTURE WORK

There are a number of directions which can be used to expand the possibilities of IMVF-LLM. Real time multimodal analytics is one such research direction which has been a priority especially in dynamic environment where data streams change at a high rate. Streaming mechanisms could be added in the future to architectures to integrate live textual commentary and video feeds as well as sensor metrics into ever-refreshing visualizations. These systems would need optimized inference pipelines - these may use lightweight model variants - to ensure cross-modal coherence and use low processing latency. Such capabilities would be the most useful in autonomous monitoring applications, such as urban traffic management and remote environmental sensing.

Another important direction of research is explainable AI visualization, which solves the problem of the lack of clarity of joint language-vision representations. Transparent output interfaces that break down decision paths the dependence of textual queries on image judgements or the dependence of numerical patterns on narrative overlays would make models much more transparent. Such explainability mechanisms are critical in a domain-sensitive setting like clinical diagnostics and market visuals associated with sentiment streams, i.e. in settings where diagnostic images are accompanied by patient histories or market visuals by sentiment streams.

Lastly, the aspects of scalability, as well as the ethical ones, should be researched systematically. To retain narrative faithfulness at scale, distributed structures that can divide multimodal data between the computational clusters are required. Visualization rhetoric mitigation of bias, privacy of sensitive image-text pairs, inbuilt auditing mechanisms to identify representational differences between demographics and sources of data should be integrated in the future system designs. To follow the indicated directions, IMVF-LLM would become a reliable and feasible analytics ecosystem.

IX. CONCLUSION

The paper has introduced IMVF-LLM, a multi-modal visualization platform that combines advanced multimodal fusion algorithms with reasoning based on large language models in order to introduce an automated system that generates meaningful visualizations on heterogeneous sources of data. Through generating contextual narratives through textual semantics and visual and structured data, the framework brings to light patterns hidden in raw data, and the process of interpreting data takes data interpretation to another level beyond the constraints of manual curation.

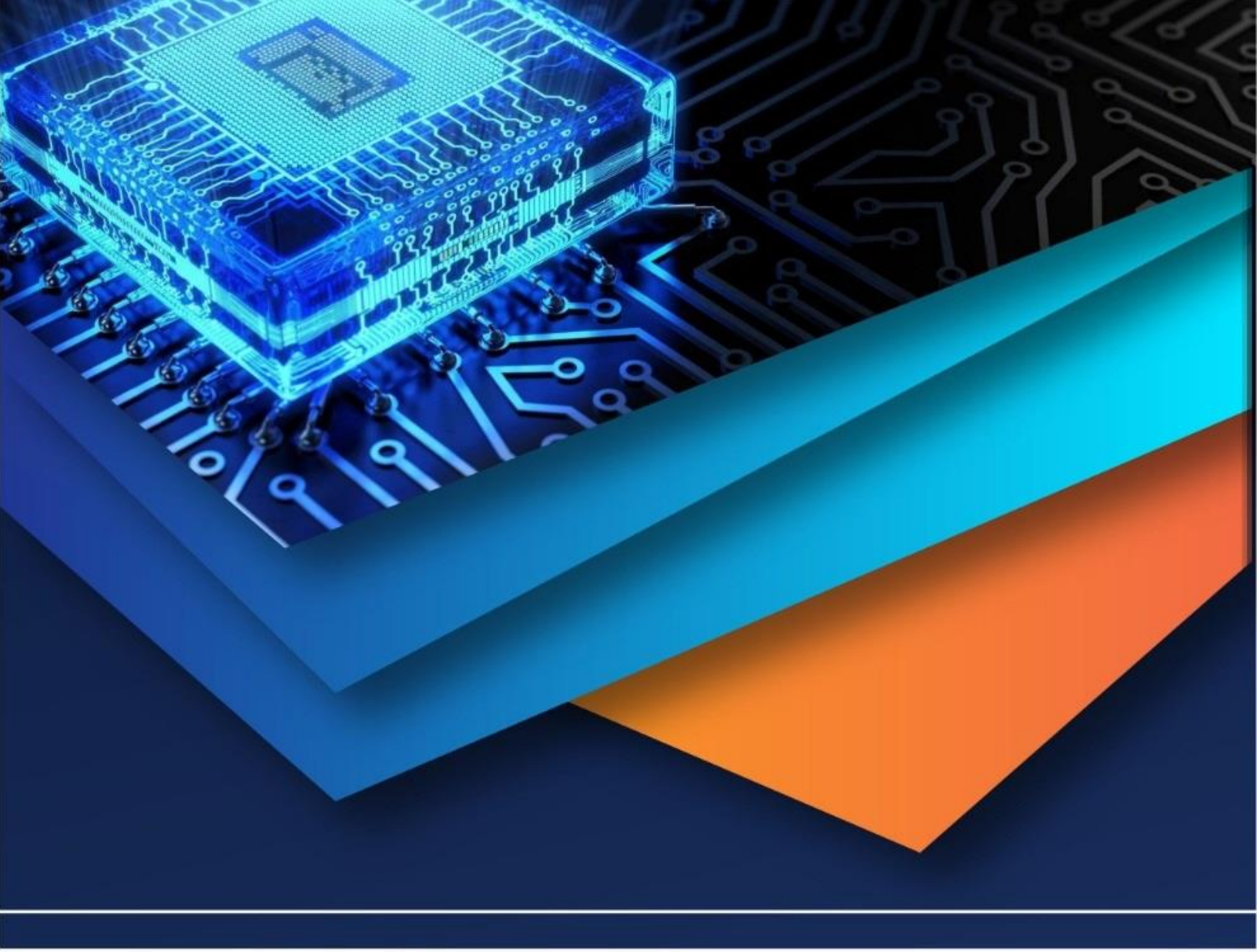
The outcome of experimental results and user studies ensure that IMVF-LLM is significantly more effective than unimodal and purely visual methods, in terms of comprehension, retention, and interpretability measures. The ability of the framework to match visualizations to the characteristics of queries, expose surface anomalies through the intuitive visualizations, and show causal relationships makes the framework a meaningful addition to scalable and human-friendly data analytics. Future research on real-time processing, explainability, and ethical robustness will further cement the use of this framework to other areas of real-world, providing tools that are useful to analysts to substantively enhance analytical reasoning of multi-faceted and complex data.

REFERENCES

[1] A. Radford et al., Learning transferable visual models using natural language supervision, in Proc. 38th Int. Conf. Mach. Learn. (ICML), PMLR, 2021, pp. 8748–8763.
[2] J.-B. Alayrac et al., Visual language model: flamingo: few-shot learning, Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 35, 2022, pp. 23716–23736.
[3] J. Li, D. Li, C. Xiong and S. Hoi, BLIP-2: Bootstrapping languageimage pre-training with frozen image encoders and large language models, in Proc. Int. Conf. Mach. Learn. (ICML), PMLR, 2023, pp. 19730–19742.

[4] D. Zhu et al., "MiniGPT-4: Vision-language understanding with enhanced large language models, arXiv preprint arXiv:2304.10592, 2023.
[5] OpenAI, "GPT-4 technical report arXiv preprint arXiv:2303.08774, 2023.
[6] E. Segel and J. Heer, Narrative visualization Telling stories with data, IEEE Trans. Vis. Comput. Graph., vol. 16, no. 6, pp. 1139–1148, Nov./Dec. 2010.
[7] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, Multimodal machine learning: A survey and a taxonomy, IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423443, Feb. 2019.
[8] J. Ngiam et al., Multimodal deep learning, in Proc. 28 th Int. Conf. Mach. Learn. (ICML), Bellevue, WA, USA, Jun. 2011, pp. 689–696.
[9] N. Srivastava and R. Salakhutdinov, Multimodal learning by deep Boltzmann machines, in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Lake Tahoe, NV, USA, 2012, pp. 2226–2234.
[10] J. Hullman and N. Diakopoulos, Visualization rhetoric: Framing effects in narrative visualization, IEEE Trans. Vis. Comput. Graph., vol. 17, no. 12, pp. 2231–2240, Dec. 2011.
[11] C. N. Knaflic, Storytelling with Data: A Data Visualization Guide to Business Professionals. Hoboken, NJ, USA: Wiley, 2015.
[12] Z. Wang, R. Li, J. Wang, F. Wu, and Y. Zhao, H.c.v.f plus future directions of visualization, IEEE Trans. Vis. Comput. Graph., vol. 26, no. 12, pp. 3400–3414, Dec. 2020.
[13] Z. Zhang et al., "Multimodal data interactive visualisation analysis, IEEE Trans. Vis. Comput. Graph., vol. 26, no. 1, pp. 802–812, Jan. 2020.
[14] V. Chandrasekaran et al., Cross-modal learning to multimodal fusion in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Shenzhen, China, 2021, pp. 16.
[15] M. Chen et al., Evaluating large language models trained on code, arXiv preprint arXiv:2107.03374, 2021.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)