

A Luminous Pen for Air-Writing: Hybrid Approach with Computer Vision and Vision–Language Models

Uma Caneta Luminosa para Escrita no Ar: Abordagem Híbrida com Visão Computacional e Modelos Visão–Linguagem

Luma T. L. de Souza^{1*}, Rafael A. D. Caldeira¹, Sérgio D. C. Leal¹, Maria Clara P. de Souza¹, Thiago M. Paixão¹, Richard J. M. G. Tello¹

Abstract: Air writing is a form of human–computer interaction that enables users to write in the air through different technological approaches, such as radio waves, wearable sensors, dedicated devices, and computer vision. In this work, we present a hybrid solution that combines the use of a luminous-tip pen with computer vision techniques to support pen detection and the air-writing process. The pen was designed and fabricated with a 3D printer and integrates an ESP32 microcontroller board, which provides Bluetooth connectivity for real-time communication with the computer. In addition to tracking movements for writing, the device can transmit specific commands to the system, such as starting or stopping the writing process and changing colors. This approach bridges the gap between wearable devices and vision-based methods, offering a practical and accessible alternative for air-writing applications. The system integrates vision–language models capable of recognizing both words and images. To determine which models were best suited for our system, we evaluated the performance of several candidates. Five participants contributed to the creation of the dataset. For drawing recognition, each participant produced images in the following classes: Tree, Moon, Cat, Heart, and Pen. For word recognition, the dataset included the following eight words: Purple, Window, Jungle, Pillow, Team, Doctor, Words, and Science. The results showed that the best-performing models achieved 76.00% accuracy with Kosmos-2 for image detection and 97.39% accuracy with Gemini 2.5 Flash for word detection, demonstrating satisfactory outcomes for the system implementation. The proposed system demonstrates the potential of combining electronic devices and advanced machine learning techniques for air-writing, with applications in education and assistive technologies.

Keywords: VLM — air writing — human-computer Interaction (HCI) — image recognition — word recognition

Resumo: A escrita no ar (air writing) é uma forma de interação humano-computador que permite escrever no ar por meio de tecnologias como sensores, dispositivos vestíveis e visão computacional. Este trabalho apresenta uma solução híbrida que combina uma caneta com ponta luminosa e técnicas de visão computacional para detectar e acompanhar o movimento da caneta durante a escrita. A caneta foi projetada em impressora 3D e utiliza um microcontrolador ESP32 com Bluetooth, permitindo comunicação em tempo real com o computador. Além do rastreamento, o dispositivo envia comandos como iniciar/parar a escrita e mudar cores, unindo as vantagens de dispositivos vestíveis e métodos baseados em visão. O sistema integra modelos de visão e linguagem (VLMs) capazes de reconhecer palavras e imagens. Foram testados vários modelos com um conjunto de dados criado por cinco participantes. Para desenhos, as classes incluíram Árvore, Lua, Gato, Coração e Caneta; e para palavras: Roxo, Janela, Selva, Travesseiro, Equipe, Doutor, Palavras e Ciência. Os melhores resultados alcançaram 76% de acurácia com o Kosmos-2 para imagens e 97,39% com o Gemini 2.5 Flash para palavras, mostrando desempenho satisfatório. O sistema demonstra o potencial da integração entre dispositivos eletrônicos e inteligência artificial para aplicações em educação e tecnologias assistivas.

Palavras-Chave: VLM — escrita no ar — interação humano-computador (IHC) — reconhecimento de Imagens — reconhecimento de palavras

¹ Instituto Federal do Espírito Santo (IFES), Brazil

*Corresponding author: lumatavares501@gmail.com

DOI: <http://dx.doi.org/10.22456/2175-2745.150973> • Received: 15/10/2025 • Accepted: 03/11/2025

CC BY-NC-ND 4.0 - This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

1. Introduction

Human–computer interaction (HCI) has traditionally relied on devices such as keyboards, mice, and, more recently, touchscreens. Emerging interaction paradigms, however, aim to move beyond these conventional interfaces, functioning as cognitive extensions that seamlessly connect users with digital environments [1]. With rapid technological advancements, new modalities of HCI have emerged, among which air-writing has gained prominence as an innovative interaction method that allows users to write in the air without relying on traditional media such as paper or blackboards.

Air writing (AW) refers to an emerging form of interaction in which gestures performed by the user in mid-air are captured and converted into letters or words displayed on digital devices [2]. This technology can also support users with various limitations, such as dyslexia, offering alternative means of interaction and communication [3].

Moreover, Air Writing methodologies can be taxonomically distinguished into diverse classifications [4]:

- **Radio Waves:** Uses RF, Wi-Fi, mmWave, and 3D ultrasonic positioning with electronic 3D pens.
- **Devices:** Involves development of wireless peripherals such as mice or 3D pens equipped with IMUs, Leap Motion Controllers, and wireless microprocessor control.
- **Wearable Sensors:** Employs inertial measurement units (IMU), motion sensors, and EEG acquisition systems.
- **Computer Vision:** Utilizes techniques like DSNT, Bi-LSTM, CNNs, signature representations with Point-Clouds, and Autoencoders.

In this work, we propose a hybrid technology for air writing that combines electronic devices with computer vision techniques. An electronic pen was developed using 3D printing and embedded with an ESP32-based microcontroller. The pen features a luminous tip, enabling reliable detection through computer vision algorithms.

Additionally, Vision-Language Models (VLMs) were incorporated into the system, providing handwriting recognition and image analysis capabilities. These models interpret written and drawn content, with accuracy evaluated based on the number of correctly recognized letters for handwriting and the ability to correctly describe or identify objects in images.

Overall, this work presents a hybrid air-writing system integrating hardware and software components. The performance of VLMs was rigorously assessed, and different models were benchmarked to determine the most effective implementation. The proposed system demonstrates the potential of combining electronic devices with advanced machine learning techniques to enable novel forms of human–computer interaction, offering enhanced accessibility and improved user experience.

2. Related Work

To support the development of this project, an analysis of published studies on air writing and related technologies was conducted.

A foundational reference for this work [5], developed an Air Writing system capable of tracking finger movements and interpreting gestures. Their approach integrated word recognition models, including HTR (Handwritten Text Recognition) [6], which achieved an accuracy of 80.90% under optimal conditions, and TrOCR [7], which reached 95% accuracy in the same context.

In a separate study, [8] introduced a CNN-based model for recognizing words written in the air. The model was trained on a dataset of 26,000 characters and featured layers for input, convolution, dense processing, and output. Dropout was applied to the first two convolutional layers as a regularization technique to reduce overfitting, enhancing the network’s generalization capabilities. The system attained a word recognition accuracy of 97.29%.

An alternative approach was proposed by [9], where smart glasses were utilized to capture RGB images. The system leveraged a modified Mask R-CNN to detect the user’s fingertip, thereby enabling real-time recognition of air-written input. In this work, the model was evaluated using the Average Pixel Error (APE), which quantifies the distance in pixels between the predicted point and the ground-truth annotation. The model achieved an APE of 10.17.

Furthermore, in [10], an air-writing system was developed based on movement trajectories. A camera is used to capture the image of a specific color — whether from a pen or the hand itself — to track the writing motion. After acquiring the images, the system employs a trained CNN model to recognize letters from A to Z, achieving an accuracy of 95%.

Moreover, in [11], the Mica Notes system was developed, which includes a device called Mica Pen — an integrated device equipped with an MTS310 accelerometer used to enable air writing. The device can be worn as a writing tool, similar to a wristwatch or bracelet. By using the accelerometer, it is possible to detect hand movements. The system is capable of recognizing eight distinct hand gestures and interpreting them based on a predefined database of letters and numbers.

While previous studies have made significant contributions to the development of air-writing technologies, several aspects distinguish the present work. Whereas studies such as [5, 9, 10] rely on finger-based gestures to perform air writing, our approach employs an electronic device that facilitates the writing process and integrates multiple research areas into a single framework, resulting in a more robust and reliable system.

Moreover, many existing systems rely on pretrained models for word and number recognition. By integrating and benchmarking multiple publicly available models, our work provides a simplified and more accessible approach for replicating and extending air-writing systems, while identifying the most effective models for both image and word recognition.

3. Methods

3.1 Pen Construction

For the development of the device, different microcontrollers were tested, including the ESP32 Mini, the standard ESP32, and an Arduino board. The final choice was the ESP32 DEVKit v1 (30 pins), as it provides both Bluetooth and Wi-Fi connectivity. Its compact form factor also simplifies integration into embedded systems, including the proposed writing tool.

The prototype consists of the ESP32 microcontroller, resistors, a 9V battery, a 10 mm LED diode (emitting white light with a bluish hue), and non-latching push buttons in four colors: blue, red, black, and green. These buttons correspond to the available color options in the writing system. The housing was produced with a 3D printer, allowing precise prototyping and easy customization.

The microcontroller communicates with the computer by sending signals to start or stop writing and to change colors. The computer used in the experiments features an Intel® Core™ i7-9700 CPU @ 3.00 GHz and 24 GB of RAM. It processes the signals with Vision-Language Models (VLMs) for handwriting and image recognition. Each domain of the system has its own dedicated functions, supported by specialized models for specific tasks. An image of the electronic stylus is shown in Figure 1.

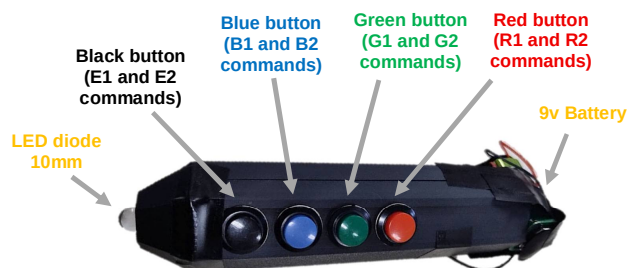


Figure 1. The figure presents a luminous pen along with its external components, including push buttons.

In addition, a webcam is employed to detect the LED signal, allowing the system to capture the light source and support the writing process. The detection operates as follows: the 10 mm diode is captured by the webcam, and the image is converted to the HSV color space. A specific value range is then defined to generate a binary mask that highlights the LED while suppressing the background. Using OpenCV, this point is tracked and translated into user trajectories, enabling air-writing recognition.

Although effective for prototyping, the method is sensitive to ambient lighting, device orientation, and occlusions, which may compromise accuracy. This process is illustrated in Figure 2, showing the system in real-time operation.

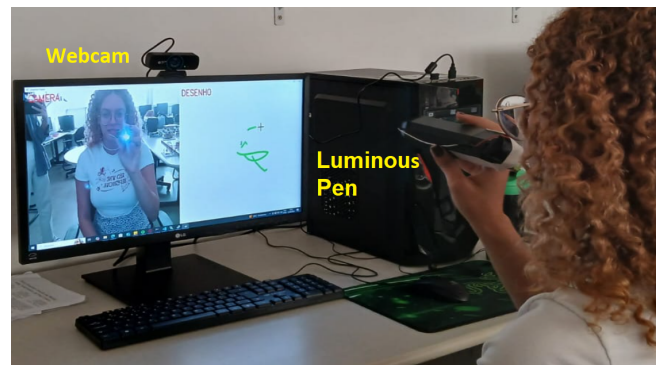


Figure 2. Real-time tests demonstrating the detection of the pen's light.

In summary, the overall system workflow can be described as follows:

- Initialization of the camera and detection of the LED tip.
- The push button on the pen sends a signal to initiate the writing process.
- Once the LED is detected, it serves as the reference point for air-writing.
- The user indicates whether the input corresponds to a drawing or a word.
- The system employs VLMs to perform recognition, processing either figures or handwriting accordingly.

Together, these components and processes demonstrate the hybrid implementation of air-writing technologies, incorporating computer vision techniques and the use of embedded devices. In addition, the system workflow, which integrates both hardware and software components, is depicted in Figure 3.

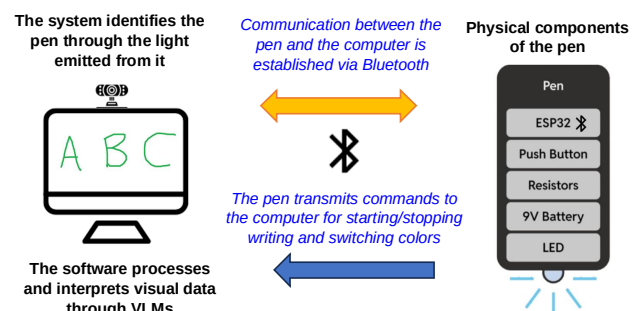


Figure 3. Illustration of the system workflow.

3.2 Text and Image Recognition

After images are captured using computer vision techniques, Vision-Language Models (VLMs) are employed for recognition tasks. Unlike conventional LLMs, VLMs jointly process and integrate visual and textual inputs to achieve a more comprehensive understanding [12]. They typically consist of two

main components: a vision encoder, which extracts high-dimensional visual features, and a language encoder, which transforms text into semantic embeddings.

In this study, VLMs were applied to both word and image recognition. For word recognition, Perplexity.ai with ChatGPT 5.0 and Gemini 2.5 Flash were used. For image recognition, VIT GPT-2 Image Captioning, Kosmos-2 [13], BLIP Base [14], GIT Base COCO, and Gemini 2.5 Flash were employed. All the models, with the exception of Plerpetixcy IA and Gemini 2.5 Flash, are available HuggingFace website¹. Furthermore, the basic version of each model—without any additional configurations—was used for the tests.

3.3 Experiment

In this work, the performance metrics were based on the models' ability to recognize words and drawings created by the users. To quantify this performance, two lists were prepared: one containing words and the other containing drawings, whose items are detailed in the Table 1. In building the dataset², five participants contributed by writing the words and then drawing the images. Each word and drawing was individually stored as an image.

Table 1. Images and words used for the dataset creation

Image Classes	Words
Tree	Purple
Moon	Window
Cat	Jungle
Heart	Pillow
Pen	Team
-	Doctor
-	Words
-	Science

Subsequently, the images containing words were processed using Perplexity.ai with the ChatGPT 5.0 model, as well as with Gemini 2.5 Flash, in order to calculate the accuracy of each model. In parallel, the images containing the drawings were submitted to five different models: VIT-GPT2 Image Captioning, GIT Base COCO, Kosmos-2, BLIP Image Captioning Base, and Gemini 2.5 Flash were also used. For each image, a description was generated by the respective model regarding its interpretation of the drawing, and based on these descriptions, the accuracy of each model was determined.

Furthermore, the ability of the VLMs to recognize images and words was analyzed, and their accuracy was calculated using the following formula:

$$\text{Accuracy} = \frac{\text{correct outputs}}{\text{inputs (letters / images)}} \times 100\% \quad (1)$$

¹<https://huggingface.com/>

²<https://bit.ly/3KRHXwb>

For the accuracy assessment of image recognition, the following evaluation criteria were adopted: if the model correctly identifies the complete description or image class, it is assigned a score of 1; if the prediction partially matches the expected output (covering approximately half of the correct answer), it is assigned a score of 0.5; and if no relevant match is achieved, the score is 0. The adopted evaluation scheme for full correctness, partial correctness, and incorrectness is illustrated in Figure 4.

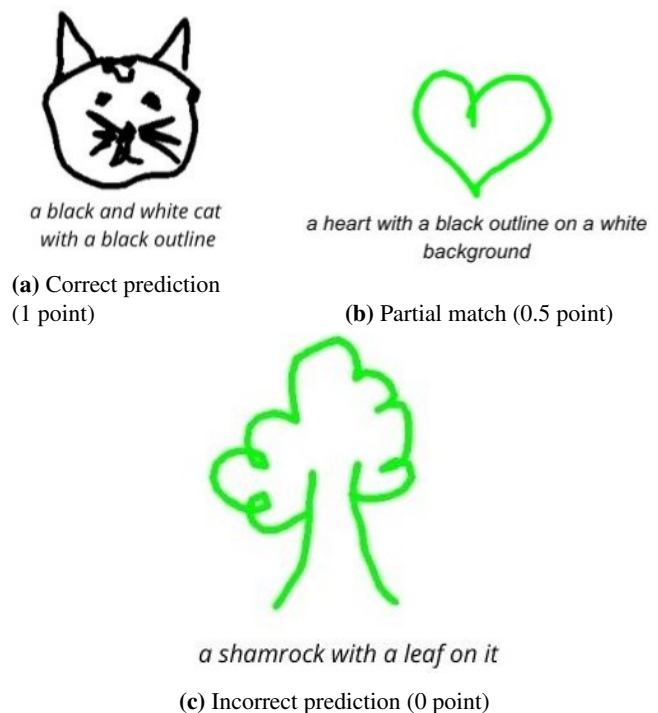


Figure 4. Illustration of the metrics implemented for scoring: correct, partial, and incorrect predictions.

Moreover, the accuracy of the word-recognition models was computed by decomposing each target word into its constituent characters (e.g., e-x-e-m-p-l-e). The predicted output generated by the models was then aligned and compared with the ground-truth sequence on a character-by-character basis. For instance, if the target word was purple (p-u-r-p-l-e), each predicted character was individually evaluated against the corresponding ground-truth character. The total number of correctly recognized characters was then divided by the overall number of characters, yielding the final accuracy measure. An illustrative example of the model output in the context of word recognition is presented in Figure 5.



Figure 5. Word and model output for word recognition.

4. Results and Discussion

The results of the handwriting experiments, presented in the Table 2, were generated using the web-based versions of the models, thus ensuring both standardization and elimination of concerns regarding token limitations. All experiments were conducted under identical input and evaluation conditions.

Table 2. Model accuracy comparison

Modelo	Accuracy
Gemini 2.5 Flash	97.39%
Perplexity.ai (using ChatGPT 5.0)	94.21%

Overall, both handwriting recognition models achieved high levels of accuracy: 94.21% for Perplexity.ai and 97.39% for Gemini 2.5 Flash. The primary distinction between the two systems lies in their handling of atypical symbols resulting from poorly formed characters. For example, Perplexity.ai occasionally misclassified ambiguous strokes as visually similar but distinct symbols, while Gemini demonstrated a more consistent ability to correctly resolve such ambiguities.

Another factor that influenced experimental outcomes was the participants' prior familiarity with the device. Most volunteers had previous experience using the luminous pen, which contributed to clearer handwriting and reduced error rates, thereby minimizing difficulties related either to device operation or the production of the writing itself. Furthermore, the image recognition results are presented in Table 3

Modelo	Acurácia
Kosmos-2 [13]	76.00%
Gemini-2.5-flash	68.00%
GIT Base COCO	50.00%
BLIP Image Captioning Base [14]	36.00%
VIT GPT2 Image Captioning	2.00%

Table 3. Performance of the models for image detection.

Overall, Gemini 2.5 Flash and Kosmos-2 achieved the highest accuracies, with 68.00% and 76.00%, respectively, demonstrating a strong ability to correctly classify the images. BLIP Image Captioning Base and GIT Base COCO showed moderate performance, with accuracies of 36.00% and 50.00%, respectively. VIT-GPT2 Image Captioning performed poorly in this task, achieving only 2.00%.

These differences in performance can be attributed to several factors, including the size and diversity of the datasets used for training. Models trained on larger and more varied datasets, such as Gemini 2.5 Flash and Kosmos-2, tend to generalize better to novel inputs like user-generated drawings and words, while those trained on smaller or more specialized datasets may perform worse. Additionally, architectural differences—such as how vision and language encoders are integrated and how outputs are generated—can affect a model's ability to handle stylized or handwritten content.

In the proposed work, the Mica Notes system was developed, which includes a device called Mica Pen — an integrated device equipped with an MTS310 accelerometer used to enable air writing. The device can be worn as a writing tool, similar to a wristwatch or bracelet. By using the accelerometer, it is possible to detect hand movements. The system is capable of recognizing eight distinct hand gestures and interpreting them based on a predefined database of letters and numbers.

5. Conclusions

In this study, we developed an integrated hardware–software system for detecting the luminous point emitted by an electronic pen. The pen was fabricated using 3D printing technology and essential electronic components, including push buttons, resistors, and the ESP32 microcontroller platform. The system supports interactive functionalities such as dynamic color switching. Pen light detection is achieved through computer vision techniques, enabling intuitive interaction while providing a reliable framework for real-time tracking and visualization of the pen's movements. This combined approach not only enhances the robustness and accuracy of air-writing recognition but also offers a straightforward and efficient method for implementing air-writing technologies in practical applications.

Furthermore, a benchmarking study was conducted on several artificial intelligence models for both image and handwriting recognition tasks, using data collected from five participants. Each participant wrote seven words and produced five drawings. The evaluation results indicated that the Kosmos-2 model achieved the highest performance in image recognition, with an accuracy of 76.00%, demonstrating its robustness and reliability in interpreting diverse and potentially ambiguous user-generated drawings. Regarding handwriting recognition, Gemini 2.5 Flash proved to be the most effective model among those tested, achieving 97.39% accuracy, which can be attributed to its superior capability to accurately recognize poorly legible characters.

Moreover, during the testing of the pen, several limiting factors were observed. Ambient lighting conditions and the distance of the pen from the camera significantly affected the detection of the luminous point, often leading to missed detections or inaccurate trajectory reconstruction. Additionally, variations in pen orientation and potential occlusions further compromised the reliability of the system in real-world usage. These limitations highlight the need for developing a more robust and resilient method for air-writing detection, potentially incorporating adaptive algorithms or hybrid approaches that combine computer vision with other sensing technologies. Such improvements would enhance the accuracy, consistency, and usability of air-writing systems in diverse and uncontrolled environments.

Despite these challenges, the results demonstrate the potential of the proposed system as a practical and accessible tool for air-writing applications. The integration of low-cost

hardware, 3D printing, and AI-based recognition not only provides a functional prototype but also establishes a foundation for future improvements in performance and usability. By addressing the identified limitations, subsequent iterations could enable more accurate real-time tracking, greater adaptability to varied environments, and enhanced user interaction. Ultimately, these advancements would broaden the applicability of air-writing systems across education, assistive technologies, and human–computer interaction, fostering more natural, intuitive, and inclusive digital experiences.

Acknowledgments

All the authors thank the Federal Institute of Espírito Santo (IFES) for their support in the development of this work. In particular, the authors Luma T. L. de Souza, Sérgio D. C. Leal, and Maria Clara P. de Souza acknowledge IFES for granting the Scientific Initiation scholarships linked to the PICTI Program (Call for Proposals – PRPPG No. 03/2024).

Author Contributions

Richard Junior Manuel and Godinez Tello and Thiago M. Paixão: The professors acted as supervisors throughout the entire project, providing guidance for both the article's writing and the construction of the pen. Maria Clara P. de Souza and Sérgio D. C. Leal: They contributed to the construction of the pen: Sérgio was responsible for the physical structure and its modeling, while Maria was responsible for the electronic components, such as the microcontroller programming.

References

- [1] YANAY, T.; SHMUELI, E. Air-writing recognition using smart-bands. *Pervasive and Mobile Computing*, v. 66, p. 101183, 2020. ISSN 1574-1192. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1574119220300602>>.
- [2] CHEN, M.; ALREGIB, G.; JUANG, B.-H. Air-writing recognition—part i: Modeling and recognition of characters, words, and connecting motions. *IEEE Transactions on Human-Machine Systems*, v. 46, n. 3, p. 403–413, 2016.
- [3] VAIDYA, V.; PRAVANTH, T.; VIJI, D. Air writing recognition application for dyslexic people. In: 2022 *International Mobile and Embedded Technology Conference (MECON)*. [S.l.: s.n.], 2022. p. 553–558.
- [4] ELSHENAWAY, A. R.; GUIRGUIS, S. K. On-air hand-drawn doodles for iot devices authentication during covid-19. *IEEE Access*, v. 9, p. 161723–161744, 2021.
- [5] BARBOSA, C. E. et al. Reconhecimento de texto para sistemas air writing: Um estudo experimental. In: SBC. *Escola Regional de Informática do Espírito Santo (ERI-ES)*. [S.l.], 2024. p. 21–30.
- [6] VLOISON, V.; XIWEI, H. *Deep Learning framework for Line-level Handwritten Text Recognition*. 2021. <https://github.com/vloison/Handwritten_Text_Recognition>.
- [7] LI, M. et al. *TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models*. 2021. <<https://github.com/microsoft/unilm/tree/master/trocr>>.
- [8] ALAM, M. S.; KWON, K.-C.; KIM, N. Trajectory-based air-writing character recognition using convolutional neural network. In: 2019 *4th International Conference on Control, Robotics and Cybernetics (CRC)*. [S.l.: s.n.], 2019. p. 86–90.
- [9] CHEN, Y.-H.; SU, P.-C.; CHIEN, F.-T. Air-writing for smart glasses by effective fingertip detection. In: 2019 *IEEE 8th Global Conference on Consumer Electronics (GCCE)*. [S.l.: s.n.], 2019. p. 381–382.
- [10] KARBHARI, J.; MUKHERJI, P. Alphabet recognition using air written trajectories. In: 2023 *International Conference on Emerging Smart Computing and Informatics (ESCI)*. [S.l.: s.n.], 2023. p. 1–6.
- [11] PAL, A. Micapen: A pen to write in air using mica notes. In: 2020 *16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. [S.l.: s.n.], 2020. p. 151–154.
- [12] GHOSH, A. et al. *Exploring the Frontier of Vision-Language Models: A Survey of Current Methodologies and Future Directions*. 2025. Disponível em: <<https://arxiv.org/abs/2404.07214>>.
- [13] PENG, Z. et al. Kosmos-2: Grounding multimodal large language models to the world. *ArXiv*, abs/2306, 2023.
- [14] LI, J. et al. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. arXiv, 2022. Disponível em: <<https://arxiv.org/abs/2201.12086>>.