ImageLingo: Text-Image Pair Generation for Language Learners

Suwon Yoon, *Undergraduate, POSTECH,* Taehyeok Ha, *Undergraduate, POSTECH,*

I. Introduction

In the rapidly evolving landscape of language learning, the integration of technology has become paramount in enhancing the effectiveness and engagement of educational methodologies. This paper introduces an innovative solution: an automated platform designed to generate contextually and visually accurate sentence-image pairs, customized to meet individual learning needs. The system leverages a synergy of cutting-edge technologies and user-centered design to offer a unique and transformative learning experience.

At the core of this platform is the GPT-4 API, a pivotal component in generating intuitive and contextually relevant sentences. The GPT-4, developed by OpenAI, stands as a sophisticated language model capable of understanding and generating text with a high degree of semantic coherence and linguistic sophistication. In this system, GPT-4 is employed to create sentences and prompts that focus on user-specified vocabulary and linguistic structures, thereby aligning with the learner's individual requirements and educational goals.

To complement the textual outputs of GPT-4, the platform utilizes an advanced image generation model, Stable Diffusion XL (SDXL)/DALL-E 3. This model is responsible for transforming the textual prompts generated by GPT-4 into visually representative images. The strength of SDXL/DALL-E 3 lies in its ability to ensure visual consistency and relevance with the corresponding text, thus creating a cohesive and meaningful learning aid.

Further enhancing the platform's capabilities is the integration of the Papago API. Renowned for its accuracy and linguistic nuance, Papago offers translations of the generated sentences into a wide array of target languages. This feature is particularly beneficial for learners seeking to understand and assimilate content in different linguistic contexts, making the platform versatile and globally applicable.

The final component of this innovative solution is its Visual User Interface (UI). The UI is designed to be user-friendly, catering to learners accessing the platform via web or mobile applications. The interface presents the tailored sentence-image pairs in an interactive and engaging format, ensuring a seamless and enriching user experience. The UI design prioritizes ease of use, accessibility, and aesthetic appeal, making language learning not only effective but also enjoyable.

In summary, this paper presents a comprehensive solution that adeptly combines advanced language processing, image generation, and translation technologies with user-centric design. The aim is to revolutionize the way language is learned, making it more intuitive, visually engaging, and tailored to individual learning preferences. This platform stands as a testament to the potential of technology in enhancing educational methodologies and fostering a deeper, more immersive learning experience.

II. Problem Identification

Our team conducted an in-depth analysis of various language learning platforms and identified the following issues:
1. Multimedia Gap: Modern language learning lacks comprehensive, accurate, and diverse multimedia resources, especially in the context of generating relevant text-image pairs.
2. Contextual Challenges: Existing solutions struggle to ensure consistency and reliability in presenting text and image pairs that are culturally and contextually accurate across various languages and dialects.
3. Diverse Language Neglect: Predominantly, resources are skewed towards widely spoken languages, sidelining less common languages and dialects, limiting accessibility and inclusivity in language learning resources.
4. Nuance Misrepresentation: The subtleties of language, including idiomatic expressions and context-specific meanings, often remain unaddressed or misrepresented due to lack of visual contextual backing.

III. System Architecture

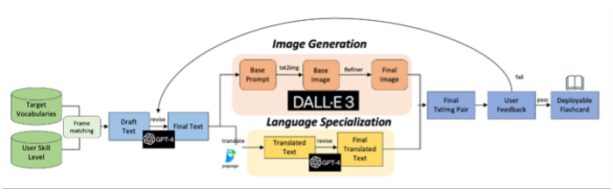

Fig. 1. 5-phased system architecture.

The architecture of this automated platform is designed to create a seamless and effective language learning experience. It integrates various components, each tailored to specific functions within the overall task flow. The process begins with user input and culminates in a unified display of the generated content.

1) User Input: The platform begins by capturing user-specified vocabulary or expressions that learners need to assimilate. This input is sourced from various materials such as textbooks, course content, or user preferences. This step is crucial as it personalizes the learning experience, ensuring that the content generated is directly relevant to the learner's current educational objectives or interests.

2) Sentence Generation: Utilizing the capabilities of GPT-4, the platform generates sentences that are contextually relevant and incorporate the specified vocabulary. GPT-4's advanced language processing abilities enable it to create sentences that are not only grammatically correct but also contextually rich, providing learners with realistic examples of how the vocabulary can be used in everyday communication.

3) Image Generation: The generated sentences or related prompts are then fed into a generative image model, such as SDXL/DALL-E 3. This step transforms the textual content into visually representative images. These images are designed to enhance comprehension and retention by providing a visual context to the newly learned vocabulary.

4) Translation: The Papago API is employed to translate the generated sentences while maintaining linguistic and contextual integrity. This feature is particularly beneficial for learners who are multilingual or are learning in a language that is not their first language. The translations help in understanding nuances and contextual meanings across different languages.

5) Unified Display: The final step involves the presentation of the original sentence, the translated text, and the corresponding image in a cohesive and interactive format via the dedicated UI. This unified display is essential for reinforcing learning, as it allows learners to see the textual and visual representations simultaneously, aiding in better comprehension and memory retention.

The architecture of this platform is meticulously designed to ensure that each component works in harmony with the others. The goal is to provide a comprehensive, intuitive, and highly engaging language learning tool that caters to the diverse needs of learners worldwide. By leveraging advanced technologies in language processing, image generation, and translation, the platform stands as a pioneering solution in the field of educational technology.

IV. Implementation

A. Target Vocabulary Acquisition and Initial Setup

The system allows users to enhance their vocabulary learning by adding target words, either manually or from course materials. To start, users upload a picture of their study material to the website. They can highlight words they want to learn, and the system recognizes and saves these in a database, creating a personalized vocabulary set.

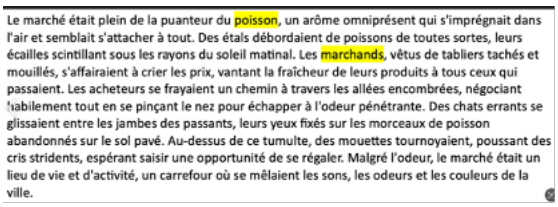
Le marché était plein de la puanteur du poisson, un arôme omniprésent qui s'imprégnait dans l'air et semblait s'attacher à tout. Des étals débordaient de poissons de toutes sortes, leurs écailles scintillant sous les rayons du soleil matinal. Les marchands, vêtus de tabliers tachés et mouillés, s'affairaient à crier les prix, vantant la fraîcheur de leurs produits à tous ceux qui passaient. Les acheteurs se frayaient un chemin à travers les allées encombrées, négociant habilement tout en se pinçant le nez pour échapper à l'odeur pénétrante. Des chats errants se glissaient entre les jambes des passants, leurs yeux fixés sur les morceaux de poisson abandonnés sur le sol pavé. Au-dessus de ce tumulte, des mouettes tournoyaient, poussant des cris stridents, espérant saisir une opportunité de se régaler. Malgré l'odeur, le marché était un lieu de vie et d'activité, un carrefour où se mêlaient les sons, les odeurs et les couleurs de la ville.

Fig. 2. Example French text.

To extract the target vocabularies, the following functions are used:

1) Image Input and Processing: The function reads the image from a specified path using OpenCV's cv2.imread. The image is then thresholded using the threshold_image function, which converts it to a binary image for easier text extraction.

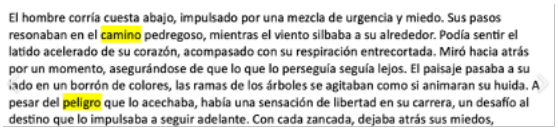
El hombre corría cuesta abajo, impulsado por una mezcla de urgencia y miedo. Sus pasos resonaban en el camino pedregoso, mientras el viento silbaba a su alrededor. Podía sentir el latido acelerado de su corazón, acompasado con su respiración entrecortada. Miró hacia atrás por un momento, asegurándose de que lo que lo perseguía seguía lejos. El paisaje pasaba a su lado en un borrón de colores, las ramas de los árboles se agitaban como si animaran su huida. A pesar del peligro que lo acechaba, había una sensación de libertad en su carrera, un desafío al destino que lo impulsaba a seguir adelante. Con cada zancada, dejaba atrás sus miedos,

Fig. 3. Example Spanish text.

2) Text Extraction Using OCR: The thresholded image is processed using Tesseract OCR (Optical Character Recognition) through pytesseract.image_to_data to extract text and its layout information.

3) Masking for Highlighted Words: The original image is masked using specific HSV (Hue, Saturation, Value) color ranges (in this case, likely targeting yellow for highlights) using the mask_image function. The masked image is then denoised using denoise_image.

4) Identifying Highlighted Words: The function find_highlighted_words is applied to the denoised mask and the OCR data to determine which words are highlighted. This is done by checking if the amount of non-zero pixels in the masked region for each word exceeds a certain threshold.

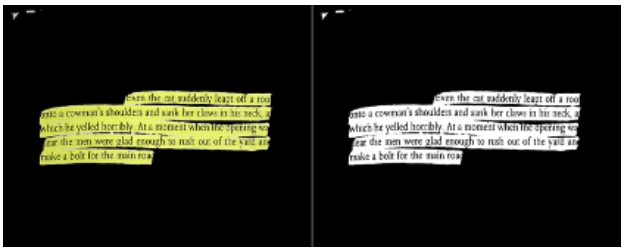

Fig. 4. Example of highlighted text being isolated for OCR.

```
def find_highlighted_words(img_mask, data_ocr, threshold_percentage=25):

    data_ocr['highlighted'] = [False] * len(data_ocr['text'])

    for i in range(len(data_ocr['text'])):
        (x, y, w, h) = (data_ocr['left'][i], data_ocr['top']
                        [i], data_ocr['width'][i], data_ocr['height'][i])
        rect_threshold = (w * h * threshold_percentage) / 100
        img_roi = img_mask[y:y+h, x:x+w]
        count = cv2.countNonZero(img_roi)
        if count > rect_threshold:
            data_ocr['highlighted'][i] = True

    return data_ocr
```

Fig. 5. Code snippet of function find_highlighted_words.

B. Text Generation

In the text generation feature, the system allows users to input words they wish to learn, either manually or by extracting them from their course materials. Once a user uploads an image and highlights the desired vocabulary, these words are sent to a function named GPT4Query. The GPT4Query function then utilizes these vocabularies to generate sentences, providing context and usage examples for each word. This approach not only reinforces the user's understanding of the vocabulary but also demonstrates its practical application in sentences.

C. Image Generation and Text Translation

The system uses DALLE-3 for image generation, creating visual aids for vocabulary learning. For translation, it detects

```python
def GPT4Query(role, instruction, context):
    messages = role
    message = "User : " + instruction
    context.extend([{"role": "user", "content": message}])
    messages = context + messages
    chat = openai.ChatCompletion.create(
        model="gpt-4", messages=messages, max_tokens=5000, temperature=0.8, top_p=0.7)
    print(f"ChatGPT: {chat.choices[0].message.content}")
    context.append(
        {"role": "assistant", "content": chat.choices[0].message.content})
    return chat.choices[0].message.content
```

Fig. 6. Code snippet of our GPT4Query function.

the target language from text in uploaded images, defaulting to English. The Papago API is used for efficient and free translation, enhancing language learning.

D. User Application

The system features an intuitive UI/UX for easy navigation, secure API communication for safe data exchange with the backend, and a comprehensive learning module to display and manage content, streamlining the user's educational experience.

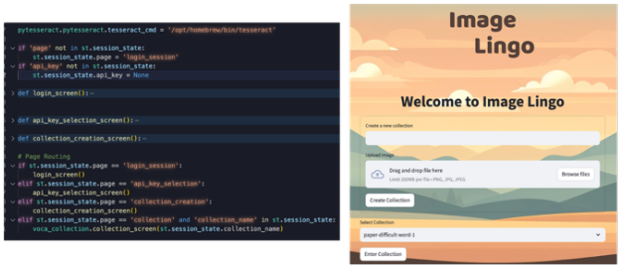

Fig. 7. Our web application UI developed using streamlit.

E. User Feedback

The system allows users to adjust the difficulty of example sentences generated by GPT-4 based on their readability level. It features two buttons: one to simplify the sentence, making it easier, and another to increase complexity, making the sentence more challenging. This customization ensures that users can tailor the learning experience to their specific skill level.

V. Results

In this section, we examine the implementation results of ImageLingo, our innovative language learning platform that integrates visual elements with language instruction. The core idea behind ImageLingo is to enhance the language learning experience by using relevant, context-rich images alongside linguistic content, thereby leveraging the cognitive benefits of visual learning.

A. Feedback and User Experience

It can be noticed right away that the incorporation of images makes the learning process more enjoyable and memorable. The visual context provided by ImageLingo aids in understanding and retaining new words and phrases, especially for abstract or complex concepts.

Fig. 8. Example of visual context aiding understanding.

B. Adaptive Learning System

One of the key features of ImageLingo is its adaptive learning system, which personalizes the sentence generation based on the user's skill level. User can ask for a harder or easier example sentence, should they feel the need for one. This system has been implemented using a fine-tuned prompt interaction based on our simulated interactions, leading to more reliable output from the language model. We believe that this personalization will lead to an increase in learning efficiency, as learners are presented with content that is optimally challenging and relevant to their current proficiency level.

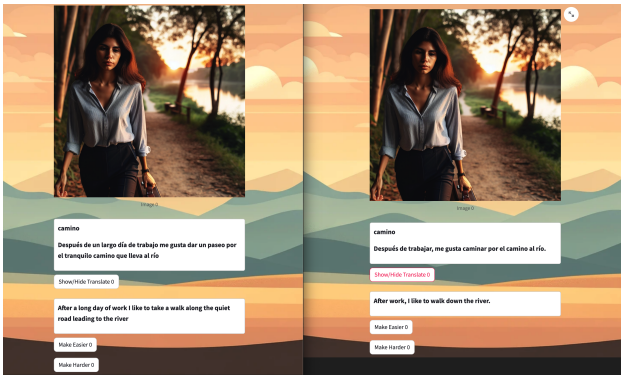

Fig. 9. Example of example sentence becoming easier after user request.

C. Global Language Support via Papago Integration

A key feature of ImageLingo's success is its integration with Papago, a versatile language translation tool. This collaboration has opened up exciting possibilities for our platform. Papago's advanced translation capabilities mean that ImageLingo

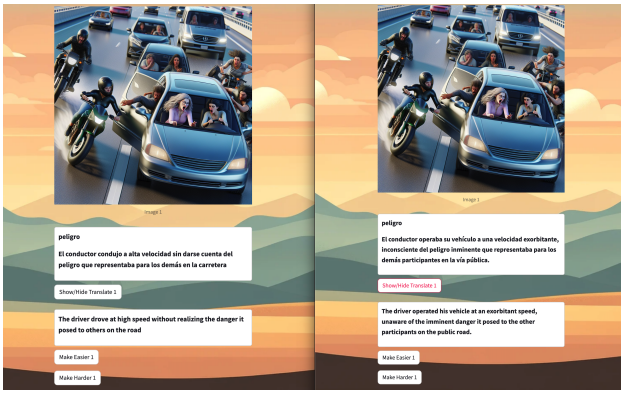

Fig. 10. Example of example sentence becoming harder after user request.

has the potential to support virtually all languages across the globe. This is a significant step towards inclusivity, allowing us to cater to a diverse user base, including those interested in less commonly taught languages. The integration with Papago not

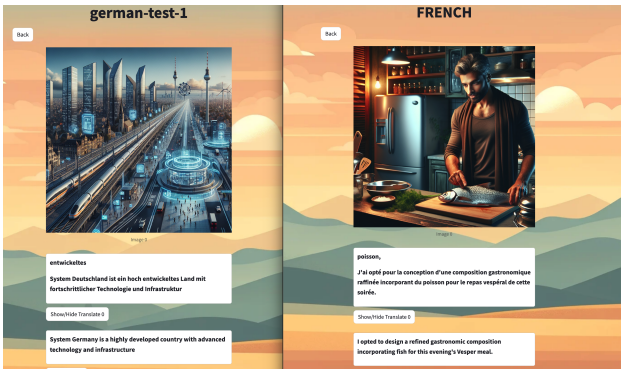

Fig. 11. Example of example sentence created in German and French.

only enhances the language variety on our platform but also ensures the accuracy and contextual relevance of translations. This is particularly crucial in language learning, where nuances and idiomatic expressions play a significant role. Our users benefit from this integration, as they can access authentic language content that accurately reflects current usage and cultural nuances.

Furthermore, this feature opens up pathways for cross-cultural exchanges and global connectivity. Learners are no longer limited to widely spoken languages; they can explore and learn languages that are specific to their interests or heritage. This aligns with our mission to make language learning accessible, engaging, and comprehensive for everyone, regardless of their linguistic background.

D. High Modularity and Scalability of ImageLingo

A defining characteristic of ImageLingo is its highly modular architecture. This design philosophy ensures that each component of the platform, such as the text recognition module, the translation module (currently integrated with Papago), and even the underlying Large Language Model (LLM), can be independently updated or replaced. This modularity is crucial for the continuous evolution and improvement of the platform.

The ability to swap out individual modules with more advanced alternatives as they are developed ensures that ImageLingo remains at the cutting edge of technology. For instance, if a more sophisticated text recognition system or a more efficient translation engine becomes available, we can seamlessly integrate these advancements into our platform without overhauling the entire system. This flexibility is essential in the rapidly evolving field of AI and language technology.

Moreover, this modular design allows us to tailor the platform to specific user needs and market demands. As new linguistic algorithms and AI models emerge, ImageLingo can easily adapt and incorporate these innovations, thereby continuously enhancing the user experience and learning efficacy. This approach not only future-proofs our platform but also aligns with our commitment to providing the most effective and up-to-date language learning tools to our users.

E. Cultural and Contextual Relevance of Images

A notable aspect of ImageLingo is the generation of culturally and contextually relevant images. This approach not only aids in language learning but also in cultural understanding. We believe this will provide a deeper appreciation and understanding of the cultural nuances of the language users are learning, which is not typically addressed in conventional language learning platforms.

VI. Discussion

In this paper, we have presented ImageLingo, a multilingual language learning platform that captures texts from real life and provides helpful example sentences. Our platform provide new insights into generative AI assited language learning. However, there are several limitations to our approach, including lack of pronunciation support, which is important part of language learning. Future work should therefore focus on improving on such limitations, potentially exploring more creative usage of LLMs.

A. Text to Speech Incorporation

Incorporating Text To Speech (TTS) technology into a language learning platform offers numerous potential benefits. Firstly, it enhances accessibility, allowing learners with visual impairments or reading difficulties to access content easily. Secondly, TTS provides learners with a practical tool for improving pronunciation and intonation, as they can hear how words and sentences are correctly articulated in the target language. This auditory exposure is crucial for developing listening skills and for acquiring the rhythm and melody of the new language. Furthermore, TTS can foster independence in learning; students can listen to text at their own pace, and repeat difficult sections as needed, facilitating self-paced learning. Additionally, for languages with non-Latin scripts or complex phonetics, TTS helps in bridging the gap between written and spoken forms, aiding in better comprehension and retention. Lastly, the integration of TTS technology can make learning more engaging and interactive, particularly for digital-savvy learners, thus increasing motivation and potentially improving learning outcomes. Incorporating TTS technlogy to ImageLingo would not be that difficult, as it would require

less than 10 lines of code to implement. However, as talked about above, its implementation would bring great benefits to users.

B. Language Learning Through Chatbots

LLMs can be effectively utilized as chatbots to assist language learners, offering a dynamic and interactive way to practice language skills. These advanced chatbots can simulate natural conversations, providing learners with a safe and judgment-free environment to practice speaking and writing. The ability to generate responses in real-time allows learners to engage in fluid dialogues, helping them to improve their fluency and conversational skills. Furthermore, LLM-based chatbots can be programmed to correct grammatical errors and suggest vocabulary improvements, offering immediate, personalized feedback that is essential for language acquisition. They can also handle a wide range of topics, enabling learners to expand their linguistic competencies in various contexts and subjects. Additionally, these chatbots can be accessible 24/7, providing learners with the flexibility to practice at their own pace and schedule. Moreover, for more advanced learners, LLM chatbots can introduce idiomatic expressions, slang, and cultural references, enriching the learning experience and providing insights into the cultural aspects of the language. This approach to language learning through conversational AI can significantly enhance engagement, motivation, and overall language proficiency. This also could be implemented along with ImageLingo, preferably sharing the database so that the generated data on the platform could be used as talking points in the chatbot. Iterative talking practices using the chatbot would enable users to practice what they have learnt from ImageLingo right away, helping users put their newly acquired knowledge into their long-term memory.

C. Drawbacks of Using LLM

Using Large Language Models in a language learning platform, while innovative, also presents several potential drawbacks. One significant concern is the risk of perpetuating and amplifying linguistic biases present in the training data, which may lead to skewed or culturally insensitive language use. LLMs may also occasionally generate inaccurate or contextually inappropriate language examples, potentially confusing learners or teaching them incorrect usage. Another issue is the lack of nuanced feedback that human instructors provide; LLMs may struggle to offer personalized, detailed critiques essential for mastering complex language aspects like idioms, slang, or subtle cultural nuances. Moreover, over-reliance on LLMs might lead to a decrease in critical thinking and problem-solving skills, as learners might become accustomed to receiving instant answers without engaging deeply with the learning material. Lastly, the integration of LLMs could potentially reduce human interaction in language learning, which is crucial for practicing conversation skills and understanding cultural contexts, thereby potentially diminishing the holistic learning experience. This is something that is currently being worked upon all around the globe, and is continually showing promising progress. So it seems probable that as time passes, these drawbacks will be resolved without having to improve ImageLingo itself.

VII. Conclusion

In conclusion, this paper has presented a groundbreaking approach to language learning, exemplifying how the integration of state-of-the-art technologies can significantly enhance the educational experience. By harnessing the capabilities of GPT-4 for generating contextually relevant sentences, coupled with the visual prowess of SDXL/DALL-E 3 for creating corresponding images, our platform offers an immersive and intuitive learning environment. The inclusion of the Papago API further elevates the platform's utility by providing accurate translations, catering to a global audience. The user-friendly Visual UI ensures that these technological advancements are accessible and enjoyable for learners of all backgrounds.

Our exploration confirms the immense potential of combining AI and user-centered design in educational tools. The platform not only simplifies the language learning process but also makes it more engaging and tailored to individual needs. As we look to the future, it is clear that the intersection of technology and education will continue to evolve, offering even more innovative solutions. This venture into the realm of AI-assisted language learning not only highlights the capabilities of current technologies but also sets the stage for future advancements that can further enhance and revolutionize the way we learn languages.

Acknowledgment

We would like to extend our heartfelt gratitude to OpenAI for granting us the opportunity to utilize their groundbreaking GPT-4 model in our research. The advanced capabilities and innovative features of GPT-4 have significantly enriched our study, allowing us to explore new frontiers in our field. This tool has not only facilitated a deeper understanding of generative AI but also inspired novel approaches and methodologies in our work. It even greatly assisted us in writing this paper.

We are also immensely grateful to Professor James Won-Ki Hong the esteemed professor of the Generative AI course. Your invaluable guidance and unwavering support have been instrumental in our journey. The opportunity to work with cutting-edge tools like GPT-4 has been a unique and enriching experience. Your enthusiasm for generative AI and commitment to fostering a learning environment that encourages exploration and innovation have profoundly impacted our academic and professional growth. Thank you for making this journey both enlightening and inspiring.

Sources

"OpenAI Platform." OpenAI Platform API Docs , platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo. Accessed 12 Dec. 2023.

"Papago Text Translation Overview." Papago Text Translation Overview, api.ncloud-docs.com/docs/en/ai-naver-papagonmt. Accessed 16 Dec. 2023.

"Dall-E 3 API." OpenAI Help Center, help.openai.com/en/articles/8555480-dall-e-3-api. Accessed 16 Dec. 2023.

Chrupała, Grzegorz, Akos Kádár, and Afra Alishahi. "Learning language through pictures." arXiv preprint arXiv:1506.03694 (2015).