

# Octopus v3: Technical Report for On-device Sub-billion Multimodal AI Agent

Wei Chen \*

Stanford University  
{weichen6}@stanford.edu

Zhiyuan Li

Stanford University  
{zhiyuan8}@stanford.edu

## Abstract

A multimodal AI agent is characterized by its ability to process and learn from various types of data, including natural language, visual, and audio inputs, to inform its actions. Despite advancements in large language models that incorporate visual data, such as GPT-4V, effectively translating image-based data into actionable outcomes for AI agents continues to be challenging. In this paper, we introduce a multimodal model that incorporates the concept of *functional token* specifically designed for AI agent applications. To ensure compatibility with edge devices, our model is optimized to a compact size of less than **1B** parameters. Like GPT-4, our model can process both English and Chinese. We demonstrate that this model is capable of operating efficiently on a wide range of edge devices, including as constrained as a Raspberry Pi. For the demo videos, try it through this link. The model weights and inference code can be applied through this link, and the model is under testing and only for research purpose now.

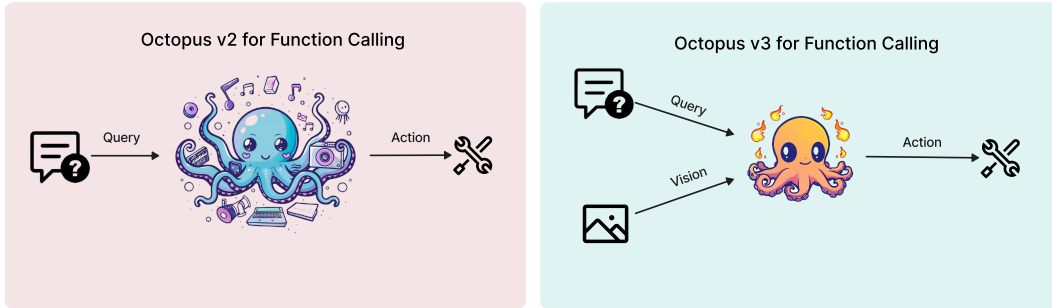


Figure 1: Sub-billion multimodal AI agent with vision

## 1 Introduction

The rapid advancement of artificial intelligence has revolutionized the way we interact with technology, giving rise to the development of sophisticated AI agents capable of performing complex tasks and making decisions based on various forms of input, including natural language [53, 44, 51, 15, 20] and visual information [9, 23]. These agents have the potential to automate a wide range of processes, from simple tasks like image recognition and language translation to more complex endeavors such as medical diagnosis [18, 34] and autonomous navigation [13, 16]. At the heart of these AI agents are multimodal language models, which enable them to understand and generate human-like responses

\*Corresponding author

by processing and integrating multiple modalities of data, such as text, images, and even audio or video.

Multimodal language models represent a significant leap forward from traditional language models, which primarily focus on processing and generating text. By incorporating visual information, these models can better understand the context and semantics of the input data, leading to more accurate and relevant outputs. For example, a multimodal language model trained on a dataset of images and their corresponding captions would be able to generate more descriptive and contextually appropriate captions for new images compared to a text-only model. The ability to process and integrate multiple modalities of data is crucial for the development of multimodal AI agents, as it allows them to perform tasks that require an understanding of both language and visual information, such as visual question answering, image-guided navigation, and multimodal sentiment analysis.

One of the key challenges in developing multimodal language models is effectively encoding visual information into a format that can be processed by the model. This is typically achieved through the use of convolutional neural networks (CNNs)[26] or transformer-based architectures [50], such as the Vision Transformer (ViT)[3]. CNNs have been widely used in computer vision tasks for their ability to extract hierarchical features from images, allowing the model to learn increasingly complex representations of the input data. On the other hand, transformer-based architectures like ViT have recently gained popularity due to their ability to capture long-range dependencies and model global context, which is particularly useful for understanding the relationships between objects in an image. These architectures enable the model to extract meaningful features from the input image and convert them into a vector representation that can be integrated with the text input.

Another approach to encoding visual information is through image tokenization, which involves breaking down the image into smaller, discrete units or tokens. This method allows the model to process the image in a manner similar to how it processes text, enabling a more seamless integration of the two modalities. Tokenized image information can be fed into the model alongside the text input, allowing the model to attend to both modalities simultaneously and generate more accurate and contextually relevant outputs. For instance, the DALL-E model developed by OpenAI uses a variant of the VQ-VAE (Vector Quantized Variational Autoencoder) [40, 35] to tokenize images, enabling the model to generate novel images based on textual descriptions.

The development of smaller, more efficient models that are capable of taking action based on a user’s query and provided image has significant implications for the future of AI agents. These models can be deployed on resource-constrained devices, such as smartphones or IoT devices, enabling a wider range of applications and use cases. By leveraging the power of multimodal language models, these smaller agents can understand and respond to user queries in a more natural and intuitive manner, while also considering the visual context provided by the user. This opens up possibilities for more engaging and personalized interactions, such as virtual assistants that can provide visual recommendations based on a user’s preferences or smart home devices that can adjust settings based on the user’s facial expressions.

Moreover, the development of multimodal AI agents has the potential to democratize AI technology, making it more accessible to a broader range of users and industries. Smaller, more efficient models can be trained on less powerful hardware, reducing the computational resources and energy consumption required for deployment. This could lead to the widespread adoption of AI agents in various domains, from healthcare and education to entertainment and e-commerce, ultimately transforming the way we live and work.

## 2 Related works

**Multimodal models** Multimodal models have gained significant attention due to their ability to process and learn from various data types, such as text, images, and audio [28, 27]. These models capture the complex interactions between different modalities and leverage their complementary information to improve performance on a wide range of tasks. Vision-Language Pre-training (VLP) models, such as ViLBERT [30], LXMERT [46], and VisualBERT [25], learn to align visual and textual features through cross-modal attention, generating rich multimodal representations. Multimodal transformer architectures, like the Multimodal Transformer (MMT) [49] and the Vision-and-Language Transformer (ViLT) [22], adapt the transformer architecture to handle multiple modalities efficiently. Researchers have also explored incorporating other modalities, such as audio and facial expressions,

into models like the Multimodal Sentiment Analysis (MSA) model [43] and the Multimodal Emotion Recognition (MER) [41] model. By leveraging complementary information from different modalities, multimodal models achieve better performance and generalization compared to unimodal approaches.

**On-device language model development** On-device language models are defined by having fewer than 7 billion parameters in our paper, since we find that it is very hard to run 13B model in an edge device even if we use quantization. Recent efforts in this area include Google’s Gemma 2B and 7B [17], Stable Diffusion’s Stable Code 3B [36], and Meta’s Llama 7B [48]. Interestingly, research from Meta suggests that smaller language models perform better with deep and thin architectures, unlike their larger counterparts. Other beneficial techniques for on-device models include embedding sharing, grouped-query attention, and immediate block-wise weight sharing, as introduced in the MobileLLM [29] paper. These findings highlight the importance of considering alternative design methods when developing smaller language models for on-device applications, as they may require different optimization approaches compared to larger models.

### 3 Methodology

This section discusses the primary techniques employed in the development of the Octopus v3 model. The two critical aspects of multimodal model development are integrating image information with text input and optimizing the model’s ability to predict actions.

#### 3.1 Encoding visual information

Numerous methods exist for encoding visual information in image processing, with embeddings from hidden layers being commonly utilized. For example, the VGG-16 [45] model’s hidden layer embeddings are used in style transfer tasks. OpenAI’s CLIP[39] model demonstrates the ability to align embeddings from both text and images, leveraging its image encoder for embedding images. More advanced techniques, such as image tokenization, are employed in approaches like ViT [28].

Our research evaluated various image encoding techniques and found the CLIP model’s method to be the most effective. Consequently, the application presented in this paper utilizes a CLIP-based model for image encoding.

#### 3.2 Functional token

Like tokenization applied to natural language and images, particular functions can be encapsulated into *functional token*[11, 12]. We introduce a training strategy for these tokens, which mirrors the techniques used by natural language models to manage unseen terms. This methodology is akin to the word2vec scheme[33], where the meaning of a token is enriched by its contextual surroundings. For instance, advanced language models might initially struggle with complex chemical terms like *PEGylation* and *Endosomal Escape*. Yet, these models are capable of acquiring such terminology through causal language modeling, especially when trained on datasets containing these terms. Similarly, functional tokens can be acquired using parallel strategies, with the Octopus v2 model serving as a robust platform for such learning processes. Our research shows that the potential for defining functional tokens is limitless, thereby allowing for the representation of any particular function as a token.

#### 3.3 Multi-stage training

To develop a high-performing multimodal AI agent, we employ a model architecture that integrates a causal language model with an image encoder. The training process for this model is structured in multiple stages. Initially, the causal language model and the image encoder are trained separately to establish a foundational baseline model. Subsequently, these components are merged, and the model undergoes alignment training to synchronize image and text processing capabilities. Following this, the training incorporates methodologies from the Octopus v2 [11] framework to facilitate the learning of *functional token*. In the final stage of training, these functional tokens, which are capable of interacting with the environment, provide feedback that is used to further refine and optimize the model. Thus, in the final stage, we use reinforcement learning, and choose another large language

model as the reward model. This iterative training approach enhances the model’s ability to process and integrate multimodal information effectively.

## 4 Model evaluation

In this section, we present the results of our model and compare them with the outcomes achieved by integrating the GPT-4V model and GPT-4 [1, 5, 38, 37]. For this comparison, we initially employ GPT-4V (*gpt-4-turbo*) to process image information. Subsequently, this extracted data is utilized within the GPT-4 framework (*gpt-4-turbo-preview*), incorporating all functional description into the context and applying few-shot learning to enhance performance. In our demonstration, we convert 10 commonly used smartphone APIs into functional tokens and evaluate their performance, as detailed in the subsequent sections.

It is important to note that while only 10 functional tokens are showcased in this paper, the model can be trained with a broader array of tokens to create a more versatile AI agent. Our findings indicate that our model, equipped with less than **1B** parameters, performs comparably to the combination of GPT-4V and GPT-4 model as multimodal AI agent for selected APIs.

Furthermore, the scalability of our model allows for the inclusion of a wide range of functional tokens, enabling the creation of highly specialized AI agents tailored to specific domains or use cases. This adaptability makes our approach particularly valuable in industries such as healthcare, finance, and customer service, where AI-driven solutions can significantly enhance efficiency and user experience.

For all the function names below, the Octopus will only output the functional token, such as `<nexa_0>`, ..., `<nexa_N>` and we replace the functional token with the corresponding function name for better demonstration. All the result below is direct generation, which doesn’t need any output parser. Octopus v3 is a single model that can handle both Chinese and English, which means that training another model especially for Chinese is not necessary.

### 4.1 Send email

Send email example	
Text and vision input	 <p>send an email to hello@nexa4ai.com with the title as AI progress. Be short, and capture the main idea;告诉hello@nexa4ai.com, 标题是AI进展</p>
Octopus v3	<pre>send_email('hello@nexa4ai.com', 'AI progress', 'The era of on device AI agent is coming! With the latest technology, businesses and individuals are embracing the power of technology in various aspects of life. The focus on innovation and efficiency in the current era is unparalleled.')&lt;nexa_end&gt; </pre>
Octopus v3 Chinese	<pre>send_email('hello@nexa4ai.com', 'AI进展', 'The era of on device AI agent is coming!')&lt;nexa_end&gt; </pre>
GPT-4V + GPT-4	<pre>send_email('hello@nexa4ai.com', 'AI progress', 'The era of on-device AI agents is coming! Advancements in AI technology are leading to more powerful and autonomous on-device AI solutions.') </pre>

Table 1: Send email examples

## 4.2 Send text message


Send text message example:	
Text and vision input	 <p>send text message to 1234567890; 发个text message给1234567890</p>
Octopus v3	<code>send_text_message('Golden bridge with cityscape in background, under a serene sky.', '1234567890')&lt;nexa_end&gt;</code>
Octopus v3 Chinese	<code>send_text_message('Golden bridge with cityscape in the background, early evening.', '1234567890')&lt;nexa_end&gt;</code>
GPT-4V + GPT-4	<code>send_text_message('Picture of the Golden Gate Bridge in San Francisco, taken at dusk with clear skies and the city skyline in the background.', "1234567890")</code>

Table 2: Send text message example

## 4.3 Google search example


Google search example	
Text and vision input	 <p>Tell me the history of this scene; 帮助我搜索它的历史</p>
Octopus v3	<code>google_search('history of the Chinese wall on the mountains')&lt;nexa_end&gt;</code>
Octopus v3 Chinese	<code>google_search('history of the Great Wall of China')&lt;nexa_end&gt;</code>
GPT-4V + GPT-4	<code>google_search('Great Wall of China history')</code>

Table 3: Google search example

#### 4.4 Amazon purchase example


Amazon purchase example	
Text and vision input	
	I want to purchase this; 我要买这个
Octopus v3	<code>amazon_purchase('White, compact, wall-mounted kitchen appliance for drying dishes, including white rice and black edges.', 'Kitchen &amp; Dining')&lt;nexa_end&gt;</code>
Octopus v3 Chinese	<code>amazon_purchase('White, compact, wall-mounted kitchen appliance for drying and cleaning, including multiple bowls and a stack of plates.', 'Kitchen &amp; Dining')&lt;nexa_end&gt;</code>
GPT-4V + GPT-4	<code>amazon_purchase('compact countertop dishwasher', 'home appliances')</code>

Table 4: Amazon purchase example

#### 4.5 Smart recycle


Smart recycle example	
Text and vision input	
	Dispose this; 回收这个
Octopus v3	<code>smart_recycle('plastic water bottles', 'plastic')&lt;nexa_end&gt;</code>
Octopus v3 Chinese	<code>smart_recycle('plastic water bottles', 'plastic')&lt;nexa_end&gt;</code>
GPT-4V + GPT-4	<code>smart_recycle('water bottles', 'plastic')</code>

Table 5: Smart recycle example

## 4.6 Lost and found


Lost and found example	
Text and vision input	
	Report my loss; 报失
Octopus v3	<code>lost_and_found('black optical computer mouse', 'Modern, black optical mouse with a streamlined design, featuring left and right click buttons and a scroll wheel, ergonomic for everyday use.')</code> <nexa_end>
Octopus v3 Chinese	<code>lost_and_found('black optical computer mouse', 'Modern, black optical mouse with a streamlined design, featuring left and right click buttons and a scroll wheel, ergonomic for everyday use.')</code> <nexa_end>
GPT-4V + GPT-4	<code>lost_and_found('optical mouse', 'A black, wired, optical mouse. It has a standard design with two buttons and a scroll wheel.')</code>

Table 6: Lost and found example

## 4.7 Interior design

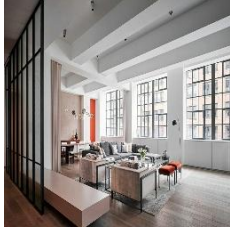
Interior example	
Text and vision input	
	Design it; 设计这个房子
Octopus v3	<code>interior_design('living room', 'modern minimalist style with neutral tones, wood furniture, and an orange ottoman for a warm touch')</code> <nexa_end>
Octopus v3 Chinese	<code>interior_design('living room', 'modern living room with white walls, black wooden beams, and an open-plan layout featuring a gray sectional sofa, black furniture, and an orange ottoman')</code> <nexa_end>
GPT-4V + GPT-4	<code>interior_design("living room", "Elegant and spacious living room with large windows, high ceilings, and a mixture of modern and mid-century furniture. Include a color scheme that complements the natural light and suggestions for decorative accessories that enhance the open, airy feel of the space.")</code>

Table 7: Interior design example



## 4.8 Instacart shopping


Instacart shopping	
Text and vision input	
	Order two for me; 帮我买两
Octopus v3	<code>instacart_shopping('pineapple', 2)&lt;nexa_end&gt;</code>
Octopus v3 Chinese	<code>instacart_shopping('pineapple', 2)&lt;nexa_end&gt;</code>
GPT-4V + GPT-4	<code>instacart_shopping('pineapple', 2)</code>

Table 8: Instacart shopping example

## 4.9 DoorDash example


DoorDash example	
Text and vision input	
	Buy and deliver to Stanford Y2E2 building; 买了之后送到同济大学
Octopus v3	<code>doordash_order('Plate of Mexican cuisine with quesadilla, guacamole, sour cream, and spicy chorizo sausage.', 'Stanford University Y2E2')&lt;nexa_end&gt;</code>
Octopus v3 Chinese	<code>doordash_order('Plate of Mexican cuisine with quesadilla, guacamole, sour cream, and spicy chorizo sausage.', '同济大学')&lt;nexa_end&gt;</code>
GPT-4V + GPT-4	<code>doordash_order('quesadilla with guacamole, sour cream, and refried beans', 'Stanford Y2E2')</code>

Table 9: DoorDash example



#### 4.10 Animal care example


DoorDash example	
Text and vision input	
	Help me take care of him; 帮我照顾他
Octopus v3	<code>animal_care('dog', 'grooming')&lt;nexa_end&gt;</code>
Octopus v3 Chinese	<code>animal_care('dog', 'grooming')&lt;nexa_end&gt;</code>
GPT-4V + GPT-4	<code>animal_care('dog', 'grooming')</code>

Table 10: Animal care example

## 5 Social impact

Enhancing the capabilities of Octopus v2, the updated model incorporates both textual and visual information, advancing from the purely text-based approach of its predecessor. This significant development allows for the simultaneous processing of visual and natural language data, paving the way for broader applications. The *functional token* introduced in Octopus v2 is adaptable and finds applications across diverse sectors, such as healthcare and automotive industries. With the addition of visual data, the *functional token* extends its potential to areas like **autonomous driving** and **robotics**. Furthermore, our multi-modal model facilitates the practical conversion of a Raspberry Pi into devices like Rabbit R1 [31] or Humane AI Pin [10], utilizing an on-device model instead of a cloud-based solution.

The *functional token* is now licensed, and we encourage developers to engage with our framework to innovate freely, in accordance with our license agreement.

## 6 Future work

In our forthcoming research, we aim to develop a training framework that accommodates additional data modalities such as audio and video. Additionally, we found that vision input may introduce considerable latency, and we are optimizing the inference speed now.

## Appendix

### A.1 Selected functions for demo

For the demonstration, we select 10 interesting functions defined on a smartphone, and we describe them below:

```
def send_text_message(message_content, phone_number):
    """
```

Send the information of the photo as a text message to a specified phone number after understanding the photo.

Parameters:

- message\_content (str): The content of the text message. This should be a concise description or information derived from the photo.
- phone\_number (str): The recipient's phone number in a format recognized by the messaging system (usually in international format).

Returns:

- bool: Returns True if the message was successfully sent, otherwise False.
- """

```
def send_email(recipient_email, subject, message_body):  
    """
```

Send the information of the photo as an email to a specified email recipient after understanding the photo.

Parameters:

- recipient\_email (str): Only one email address of the recipient.
- subject (str): The subject line of the email, should be concise.
- message\_body (str): The main content of the email.

Returns:

- bool: True if the email was successfully sent, False otherwise.
- """

```
def google_search(search_query):  
    """
```

Performs a Google search and retrieves search results based on understanding a photo.

Parameters:

- search\_query (str): The query string to search for.

Returns:

- str: Google search result.
- """

```
def amazon_purchase(product_description, product_category):  
    """
```

Initiates a purchase on Amazon for a product based on its description and category, which are derived from analyzing a photo of the product.

Parameters:

- product\_description (str): A concise description of the product to be purchased, as interpreted from the photo.
- product\_category (str): The category under which the product is listed on Amazon.

Returns:

```

- str: A message detailing the purchase outcome, including the success
status, order ID, and total price if successful, or an error message if
the purchase fails.
"""

def smart_recycle(item_name, item_category):
    """
    Provide recycling instructions for a specified item based on its
    category after understanding a photo of the item.

    Parameters:
    - item_name (str): The name of the item to be recycled.
    - item_category (str): The category of the item, must choose one from "
    electronics", "plastic", "paper".

    Returns:
    - str: Detailed instructions on how to recycle the item properly.
    """

def lost_and_found(item_name, message_content):
    """
    Report a lost item and search for it based on understanding a photo of
    the item.

    Parameters:
    - item_name (str): The name of the lost item.
    - message_content (str): The description of the lost item.

    Returns:
    - str: Instructions on how to proceed with the lost and found process.
    """

def interior_design(room_type, design_request):
    """
    Generates customized interior design suggestions for a specified room
    type based on a given design request.

    Parameters:
    - room_type (str): The type of room to be designed, such as "living
    room", "bedroom", or "office".
    - design_request (str): A specific, concise request for the interior
    design.

    Returns:
    - str: A comprehensive interior design proposal that includes
    recommendations for furniture placement, color schemes, and decorative
    accessories, tailored to the specified room type and design request.
    """

def instacart_shopping(product_name, number_of_order):
    """
    Place an order for a specified quantity of a product using the
    Instacart shopping service, the product information is from the photo.

    Parameters:

```

```

- product_name (str): The name of the product to be ordered, such as "
apples", "bananas", or "pasta".
- number_of_order (int): The quantity of the product to be ordered.

Returns:
- str: Confirmation details that include the order status, total cost,
and expected delivery time. If the order cannot be placed, an error
message will be provided.
"""

def doordash_order(food_description, address):
    """
    Places an order for delivery of a specified food item to a given
    address using the DoorDash service, the food description is from the
    photo.

    Parameters:
    - food_description (str): A description of the food item to be ordered,
    including any specific preparation requests or dietary considerations.
    - address (str): The delivery address where the food should be sent.
    Example addresses might include "123 Maple St, Springfield, IL" or "456
    Elm Ave, Apt 7, Dallas, TX".

    Returns:
    - bool: True if the order was successfully placed, False otherwise.
    """

def animal_care(animal_type, care_type):
    """
    Provides care instructions for a specific type of animal based on the
    care needed and frequency, derived from understanding a photo of the
    animal.

    Parameters:
    - animal_type (str): The type of animal, for example, "dog", "cat".
    - care_type (str): The type of care required, must choose from "feeding
    ", "grooming", "exercise".

    Returns:
    - str: Detailed care instructions tailored to the specific animal type,
    care needed, and frequency.
    """

```

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, 2022.
- [3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.
- [4] Cristian Bodnar, Adrian Li, Karol Hausman, Peter Pastor, and Mrinal Kalakrishnan. Quantile qt-opt for risk-aware vision-based robotic grasping. *arXiv preprint arXiv:1910.02787*, 2019.

- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [6] Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. Generative ai at work. Technical report, National Bureau of Economic Research, 2023.
- [7] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.
- [8] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. *arXiv preprint arXiv:2004.05155*, 2020.
- [9] Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. Evaluating visual conversational agents via cooperative human-ai games. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 5, pages 2–10, 2017.
- [10] Imran Chaudhri. Humane ai, 2024. URL <https://humane.com/>. Accessed on March 31, 2024.
- [11] Wei Chen and Zhiyuan Li. Octopus v2: On-device language model for super agent. *arXiv preprint arXiv:2404.01744*, 2024.
- [12] Wei Chen, Zhiyuan Li, and Mingyuan Ma. Octopus: On-device language model for function calling of software apis. *arXiv preprint arXiv:2404.01549*, 2024.
- [13] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979, 2024.
- [14] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2018.
- [15] Xin Luna Dong, Seungwhan Moon, Yifan Ethan Xu, Kshitiz Malik, and Zhou Yu. Towards next-generation intelligent assistants leveraging llm techniques. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5792–5793, 2023.
- [16] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 910–919, 2024.
- [17] Gemma Team, Google DeepMind. Gemma: Open models based on gemini research and technology, 2023. URL <https://goo.gle/GemmaReport>.
- [18] Stephen Gilbert, Hugh Harvey, Tom Melvin, Erik Vollebregt, and Paul Wicks. Large language model ai chatbots require approval as medical devices. *Nature Medicine*, 29(10):2396–2398, 2023.
- [19] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.
- [20] Allyson I Hauptman, Beau G Schelble, Nathan J McNeese, and Kapil Chalil Madathil. Adapt and overcome: Perceptions of adaptive autonomous agents for human-ai teaming. *Computers in Human Behavior*, 138: 107451, 2023.
- [21] Abhishek Kadian, Joanne Truong, Aaron Gokaslan, Alexander Clegg, Erik Wijmans, Stefan Lee, Manolis Savva, Sonia Chernova, and Dhruv Batra. Sim2real predictivity: Does evaluation in simulation predict real-world performance? *IEEE Robotics and Automation Letters*, 5(4):6670–6677, 2020.
- [22] Wonjae Kim, Bokyoung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.
- [23] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.
- [24] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

- [25] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
- [26] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.
- [27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- [28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [29] Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, et al. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. *arXiv preprint arXiv:2402.14905*, 2024.
- [30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.
- [31] Jesse Lyu. Rabbit r1, 2024. URL <https://www.rabbit.tech/>. Accessed on March 31, 2024.
- [32] Timothy R McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N Halgamuge. From google gemini to openai q\*(q-star): A survey of reshaping the generative artificial intelligence (ai) research landscape. *arXiv preprint arXiv:2312.10868*, 2023.
- [33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [34] Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. A study of generative large language model for medical research and healthcare. *NPJ Digital Medicine*, 6(1):210, 2023.
- [35] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10775–10784, 2021.
- [36] Nikhil Pinnaparaju, Reshith Adithyan, Duy Phung, Jonathan Tow, James Baicoianu, and Nathan Cooper. Stable code 3b, 2023. URL <https://huggingface.co/stabilityai/stable-code-3b>.
- [37] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.
- [38] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [40] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.
- [41] Nicu Sebe, Ira Cohen, and Thomas S Huang. Multimodal emotion recognition. In *Handbook of pattern recognition and computer vision*, pages 387–409. World Scientific, 2005.
- [42] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020.
- [43] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017.
- [44] Venkat Krishna Srinivasan, Zhen Dong, Banghua Zhu, Brian Yu, Damon Mosk-Aoyama, Kurt Keutzer, Jiantao Jiao, and Jian Zhang. Nexusraven: a commercially-permissive language model for function calling. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

- [45] Srikanth Tammina. Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *International Journal of Scientific and Research Publications (IJSRP)*, 9(10):143–150, 2019.
- [46] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
- [47] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [48] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [49] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.
- [50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [51] Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. *arXiv preprint arXiv:2401.16158*, 2024.
- [52] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*, 2023.
- [53] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.