

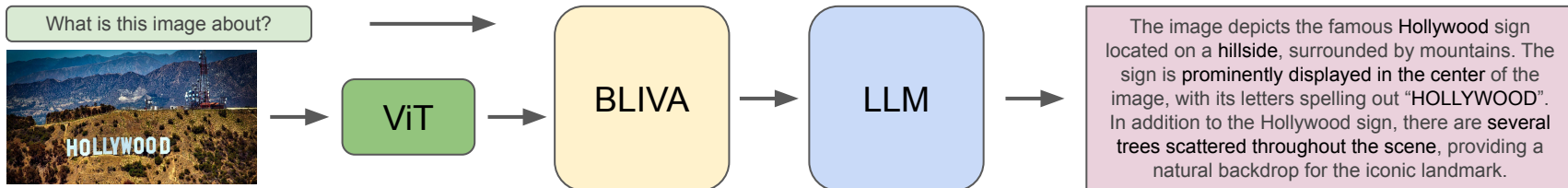
BLIVA: A Simple Multimodal LLM for Better Handling of Text-rich Visual Questions

Wenbo Hu^{*1}, Yifan Xu^{*2}, Yi Li¹, Weiyue Li¹, Zeyuan Chen¹, Zhuowen Tu¹

¹UC San Diego, ²Coinbase Global, Inc

^{*}Equal Contribution

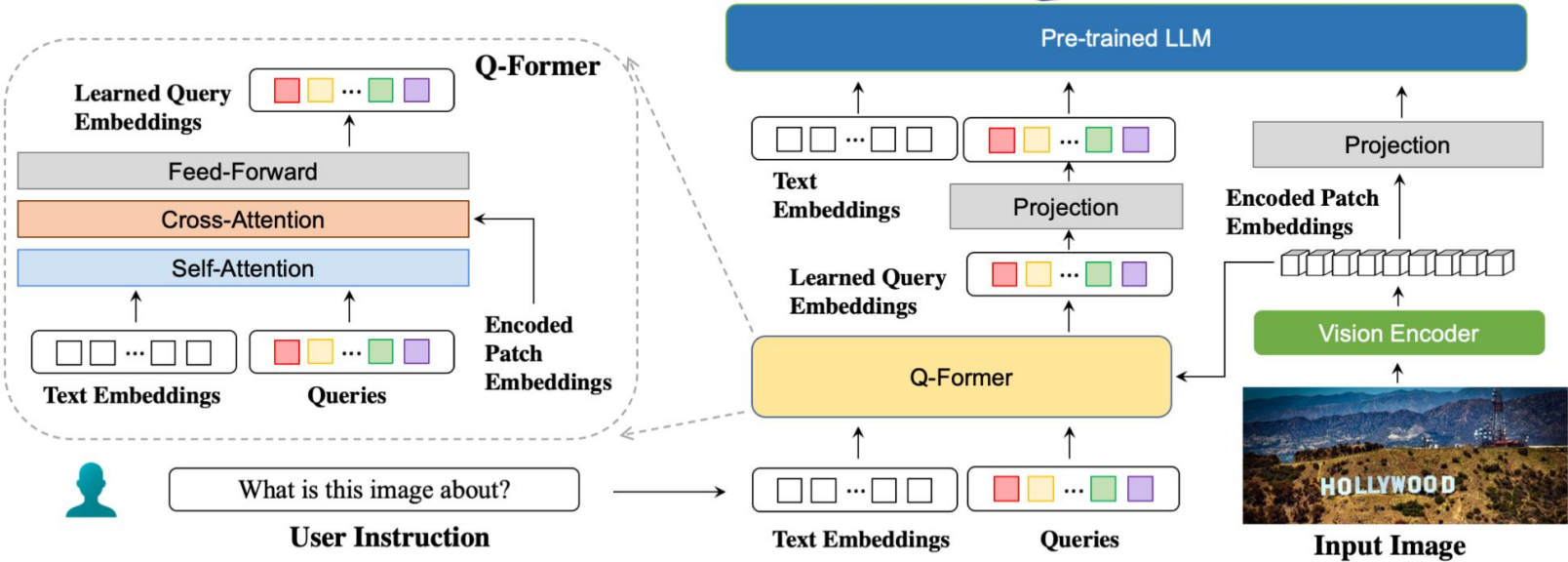
BLIVA produces high-quality **image captions**, responds **visual questions**, and extracts **textual content**.



BLIVA's Architecture:



The image depicts the famous **Hollywood** sign located on a **hillside**, surrounded by mountains. The sign is prominently **displayed in the center of the image**, with its letters spelling out "**HOLLYWOOD**." In addition to the Hollywood sign, there are **several trees scattered throughout the scene**, providing a natural backdrop for the iconic landmark.



Visual Captioning

What does it do?: It generates textual descriptions for images or video

Example Use Cases:

- **Accessibility:** Makes visual content accessible to visually impaired users by providing a textual descriptions.
- **Search Engine Optimization:** Textual captions make it easier for search engines to understand and index visual content, thus making them more discoverable.
- **User Engagement:** For social media platforms, visual captioning can increase engagement by providing more context to images and videos.
- **Archiving and Data Management:** Captioning makes it easier to categorize, sort, and retrieve visual data in large databases.

Benefits:

- **Inclusivity:** Makes visual content accessible to a larger and more diverse audience, including those with visual impairments.
- **Enhanced Understanding:** Provides additional context that might not be immediately evident from the visual content alone.
- **Efficiency:** Automated captioning can process large volumes of visual data quickly, saving time and manual effort.
- **Improved Discoverability:** Makes it easier for users and algorithms to find relevant content based on textual queries.

Demo: Visual Captioning

Industry: Food

Use Case: Write attractive descriptive food captions along with ingredients.



In the image, a person is pouring lime juice over several tacos on a wooden cutting board. The tacos are topped with various ingredients, such as **chickpeas**, **tomatoes**, and **cilantro**. The **lime juice** adds a tangy and refreshing flavor to the dish, complementing the other ingredients and making the tacos even more delicious.

The image features a delicious-looking plate of French toast topped with **sliced bananas** and **blueberries**. The French toast is placed on top of the plate, and the bananas and blueberries are scattered around it, creating a visually appealing presentation.



Demo: Visual Captioning

Industry: Sports

Use Case: Write lines for sport games.



In the image, a group of young **football** players are engaged in a **heated game** on the field. The players are wearing **helmets** and **jerseys**, indicating that they are actively participating in the game. The players are scattered around the field, with some being **tackled** and others trying to maintain possession of the ball. The game appears to be quite **intense**, as the players are pushing and shoving each other in their efforts to gain control of the ball or prevent their opponents from doing so.

Overall, the scene captures the **excitement** and **energy** of a youth football game, showcasing the **athleticism** and **competitive spirit** of the players involved.

Demo: Visual Captioning

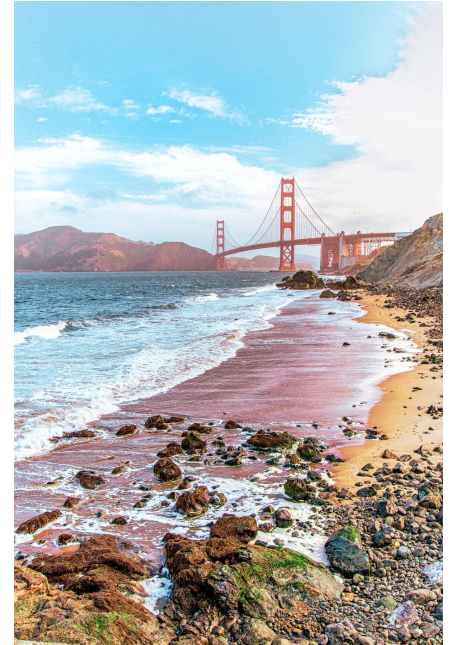
Industry: Tourism

Use Case: Write a travel introduction for a place.



The park in the image features a **large pond** with **lily pads** floating on its surface. **Palm trees** surround the pond, adding to the serene and picturesque atmosphere of the park. There is also a **clock tower** visible in the background, providing a **sense of structure and history** to the scene. Overall, the park appears to be well-maintained and beautifully landscaped, making it an ideal location for **relaxation and leisure activities**.

The place in the image is a beach with a **rocky shoreline** and a body of water next to it. There is a large bridge in the background, which is likely the **Golden Gate Bridge** in San Francisco, California. The presence of the bridge adds to the scenic beauty of the location, making it an attractive spot for **tourists and locals alike**. The combination of the rocky shoreline, the body of water, and the iconic bridge creates a **unique and picturesque setting** that showcases the natural beauty of the area.



Visual Question Answering

What does it do?: Ask a question about an image, and the computer figures it out and gives you an answer.

Example Use Cases:

- **Retail and E-commerce:** Customers could ask questions like "Does this shirt come in red?" while viewing an online product, and VQA can automatically respond.
- **Education:** Teachers and students can use VQA to make learning more interactive. For instance, "Which country does this image indicate in a geography class."
- **Tourism and Museums:** Visitors could use an app to ask questions about art pieces or historical objects they're looking at, like "Who is the artist?" or "What period is this from?"

Benefits:

- **Enhanced User Experience:** VQA allows for a more interactive and engaging user experience compared to traditional UI paradigms.
- **Increased Efficiency:** By directly asking questions, users can more efficiently find what they're looking for without navigating through menus or reading extensive documentation.
- **Reduced Cognitive Load:** Users don't have to interpret the visual data themselves but can rely on the VQA system for accurate and quick answers

Demo: Visual Question Answering

Use Case: Answer user's questions based on images.



Q: What is the man doing?

A: The man is standing on a **rocky outcropping**, looking out over a **beautiful lake** or body of water. He appears to be taking in the scenery and enjoying the view.

Q: What activity he is engaging in?

A: hiking

Q: What danger to be aware of when visiting this place ?

A: The danger to be aware of when visiting this place is the risk of **slipping** or **falling** while hiking on the rocky terrain. In the image, a man is seen standing on a rocky outcropping with a backpack, overlooking a large body of water. The **uneven** and potentially **slippery surface** can lead to accidents or **injuries** if proper precautions are not taken. Hikers should be mindful of their footing, wear **appropriate footwear**, and take necessary **safety measures** to ensure a safe and enjoyable experience in this scenic location.

Demo: Visual Question Answering

Use Case: Answer user's questions based on images.



Q: What is the name of the shop?

A: Samovar Tea House.

Q: What are the two people doing?

A: The two people are sitting at a table in front of the tea house, enjoying their drinks and engaging in conversation.

Q: What does the surrounding of the shop look like?

A: The shop is located on a city street, which suggests that it is situated in a bustling urban area.

Visual Text Understanding

What does it do?: It analyzes and interprets text that is embedded in visual content, such as images or videos.

Example Use Cases:

- **Document Analysis:** Extracting and understanding text from scanned documents, even when they include embedded images, charts, or other visual elements.
- **Retail and Inventory Management:** Automatically scanning and interpreting labels, price tags, and product details.
- **Healthcare:** Extracting and interpreting textual information from medical scans or images.

Benefits:

- **Contextual Understanding:** Providing a richer, more contextual understanding of visual data by integrating the information carried by embedded text.
- **Automation:** It can automate various tasks that would be difficult or tedious for humans to perform, such as scanning large datasets of visual content for specific textual information.
- **Enhanced Search:** Textual cues can be used to improve search functionality for visual databases.

Demo: Visual Text Understanding

Task: Image Caption

Use Case: Extract both image and text information.



In the image, there is a small wooden table placed next to a bed. On the table, there is a book titled "Milk and Honey" by Rupie Kaur.

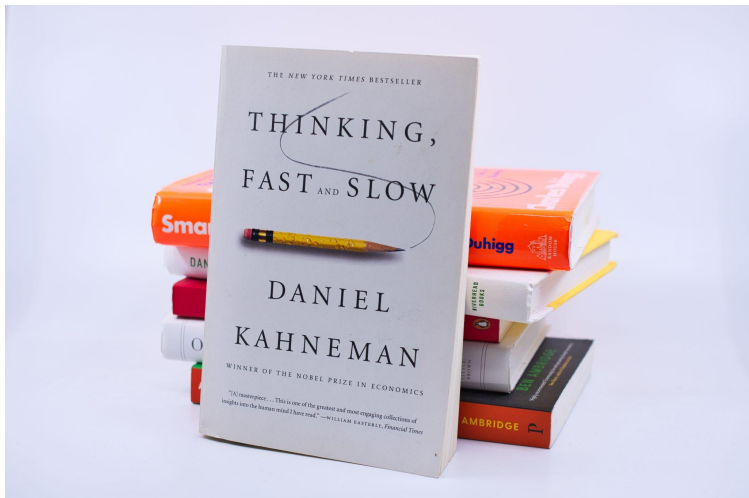
In the image, there is a book titled "The Kinfolk Table" placed on top of a blanket. The book appears to be a cookbook, and it is accompanied by a hat and a pair of scissors. There are also some yellow tulips next to the book, adding a touch of color to the scene.



Demo: Visual Text Understanding

Task: Visual Question Answering

Use Case: Extract both image and text information.



Q: What is the title of this book?

A: Thinking, Fast and Slow

Q: Who is the author of the book ?

A: Daniel Kahneman

Q: What does the book cover have?

A: The book cover of "Thinking, Fast and Slow" has a yellow pencil on it.