VisionMate: AI-Powered Image Captioning Web Application

Master's Defense By Sai Anoushka Kokku Advisor: Dr. Chris Pollett Committee: Dr. Robert Chun Committee: Dr. Thomas Austin

Outline

Introduction Background Model Research and Experimentation System Architecture and Implementation Deployment **Evolution of VisionMate** Testing, Evaluation and Results **Conclusion and Limitations** Future Work References

Introduction

Problem Statement:

- Visually impaired users cannot interpret visual scenes
- Most daily tasks rely on image-based information
- Need: A fast, simple way to describe images aloud using Al

Project Goals:

- Real-time image captioning system
- Should run on both desktop and mobile browsers
- Must give spoken feedback no visual-only results
- Lightweight, free to deploy, no installation required
- Leverage transformer-based models (via API, not locally)

Background: Related Work

Existing Assistive Tools:

- Seeing Al (Microsoft): Describes people, text, and scenes via mobile app
- **Be My Eyes:** Connects users with volunteers to describe visuals via video call

Both apps require installation and run only on mobile platforms No real-time web-based alternative for spoken image captions

Background: Related Work

Seeing Al output





ceiling and a row of overhead lights running parallel to the ceiling. The room features a patterned carpet floor and several pieces of furniture. There are four tall chairs with white backs and seats, positioned around two circular tables with metal bases. These chairs and tables are arranged in the center of the room, with two chairs on each side of the tables.

To the right, near the window, there is a person seated on a patterned armchair. The armchair has a red and orange circular pattern. The windows are large and rectangular, allowing natural light to enter the room. The walls are painted in a neutral tone, and the ceiling has a triangular shape with a circular vent or fixture at the apex. The room extends into a hallway with doars visible in the backaround.



Background: Image Captioning Overview

- Image captioning = understanding an image + generating a sentence
- Combines computer vision (object recognition) + language generation
- Requires large datasets like COCO (image + caption pairs)
- Uses encoder-decoder architecture (e.g., CNN + RNN or Transformer-based)

Background: Vision-Language Models

- Early models used CNNs + RNNs (slow, sequential, low accuracy)
- Transformers process all inputs in parallel (faster, better long-range understanding)
- Vision-Language models = vision encoder + language decoder
- Examples:

BLIP: Efficient with limited data

GIT: Generative transformer, fluent sentence output **VIT-GPT2**: Combines VIT and GPT2 for richer generation

BLIP-base vs BLIP-large – Architectural Differences

BLIP (Bootstrapped Language-Image Pretraining) uses 3 key training objectives:

- ITC: Image-Text Contrastive alignment
- ITM: Image-Text Matching for semantic relevance
- LM: Language Modeling for caption generation

BLIP-base:

- Fewer parameters (~138M)
- Faster inference, smaller model size
- Suitable for devices with limited resources **BLIP-large:**
- More layers (~370M+ parameters)
- Slower inference, high memory usage
- Generates more descriptive captions

BLIP Comparison – Performance & Output

Metric	BLIP-base	BLIP-large
Inference Time	5s to 8s	13s to 16s
Avg. Caption Length	8–12 words	11–15 words
Resource Requirement	~3.6 GB RAM (Colab)	~7.8 GB RAM
Output Style	Concise	Detailed + Natural
Suitability	Best for real-time use	Slower, better for offline use

BLIP Comparison – Performance & Output



Base: a train track with trees and bushes in the back**garged**there is a train track that is surrounded by trees and bushes (Time: 5.30s) (Time: 13.14s)



Original Image

Why Switch to GIT?

• BLIP-base generated **inaccurate captions** in some cases

Example: "Girl at a table with food" – no food or table was present

- GIT models (from Microsoft) use generative training (free-form captioning)
- Designed to mimic natural image-to-text generation
- Pretrained on COCO & Conceptual Captions datasets
- Better generalization and fewer hallucinations



GIT-base vs GIT-large – Key Differences

Metric	GIT-base	GIT-large
Inference Time	5–8 seconds	35–50 seconds
Caption Detail	Moderate, accurate	Highly detailed (e.g., gestures)
Memory Use	~4.2 GB	~9+ GB
Output Example	"A girl sitting in a library"	"A girl wearing a green T-shirt and a black jacket sitting in a Library"
Suitability	Balanced real-time use	Not suitable for web apps

Git-Base Examples



00



Close Camera

* Preview:

Dinvert More filters 40,000 ms 50.000 ms 60.000 ms 70.000 mt Name Iocalhost 299.8 443 ms Dundle.js 302 B 10 ms (index):2 at ws ebsock... react refresh:6 0B Peoding favicon.ico x-icon 363 B 21 ms D manifest.json 363 B 23 ms logo192.png 364 B 3 ms Beep-07.way 403 Other 213 8 123 ms media # blob:http://localhost:3000/94e63... 200 react-dom-client.de 08 0 ms ipeg () caption/ App.is:32 252 B 47.27 s 200 fetch

🖟 🗖 Elements Console Sources Network Performance >> 🛛 🛛 😨 🗄 🗙

🝸 Q. 🗌 Preserve log 🗌 Disable cache No throttling 🔹 😪 🙏

G & D | G :

1

Caption: a girl sitting in a library with a chair.

9 requests 2.2 kB transferred 1.3 MB resources Finish: 1.1 min DOMContentLoaded: 2.34 s Loa

		C Topic: Midtern	×	CS252-Dav13 X	CS252-D	av12-	×	Vite + React	×	React App	a post-completie x	D Post-Completi X	one SJSU	×	+	
•	•	Topoc. Miluterin	^	COTOT-Dakio. V	0 03202-0	3412	^	V VILE + REGUL	<u> </u>	Meace whith A	Ch host-combien v	Post-Complet A	Cine.3030	^		

 $\leftarrow \rightarrow$ C (D) localhost:3000



Capture Image

* Preview:



Caption: a girl with her hands on her head

R LD Elements Console Sc	ources N	etwork F	Performance >>	0 24 8	3 1 ×	
● Ø ▼ Q □ Preserve log	Disat	le cache	No throttling + 🥱	1 ± ±	۲	
Y Filter	nvert Mi	ore filters				
All Fetch/XHR Doc CSS JS F	ont Img	Media M	anifest WS Wasm	Other		
200,000 ms 400,000 ms	600.000 ms	800.0	1,000,000 ms	1,200.0	00 ms	
		-			and the second	
		-		-		
Name	Status	Type	Initiator	Size	Time	
(i) caption/	200	fetch	App.is:32	249 8	35.25	
Deep-07.way	403	media	Other	213 B	114 m	
blob:http://ocalhost:3000/ac448f	200	ipeg	react-dom-client.de	0.8	1 m	
()) caption/	200	fetch	App, is:32	252 B	25.02	
S beep-07.wav	403	media	Other	213 B	110 m	
blob:http://localhost:3000/8c91b9	200	jpeg	react-dom-client.de	08	2.m	
() caption/	200	fetch	App.is:32	261 B	21.57	
O beep-07.wav	403	media	Other	213 B	111 m	
blob:http://localhost:3000/ee873a	200	ipeg	react-dom-client.de	08	2 m	
()) caption/	200	fetch	App,is:32	257 B	25.69	
S beep-07.wav	403	media	Other	213 B	139 n	
# blob:http://localhost:3000/b75d5	200	jpeg	react-dom-client.de	08	2 m	
B beep-07.wav	403	media	Other	213 B	153 m	
# blob:http://localhost:3000/3fdaf6	200	jpeg	react-dom-client.de	08	1 m	
0) caption/	200	fetch	App.is:32	272 B	12.31	
is blob:http://localhost:3000/945bc	200	jpeg	react-dom-client.de	08	3 m	
S beep-07.wav	403	media	Other	213 B	131 п	
63 caption/	200	fetch	App.is:32	235 B	5.55	
S beep-07.wav	403	media	Other	213 B	109 m	
at blob:http://localhost:3000/7977bd	200	jpeg	react-dom-client.de	08	0 m	
0) caption/	200	fetch	App.is:32	253 B	9.94	
beep-07.wav	403	media	Other	213 B	119 m	
blob:http://iocalhost:3000/58e44	200	ipeg	react-dom-client.de	08	1 m	
() caption/	200	fetch	App.is:32	237 B	6.42	
S beep-07.wav	403	media	Other	213 B	108 m	
blob:http://localhost:3000/593bf0	200	ipeg	react-dom-client.de	08	0 m	
(3) caption/	200	fetch	App.is:32	244 8	8.18	

Git-Large Examples



Caption: a woman making a heart with her hands.

Invert More filters • All Fetch/XHR Doc CSS JS Font Img Media Manifest WS Wasm Other 20,000 ms 40,000 ms 60,000 ms 80,000 ms 100,000 ms 120,000 ms 140 000 ms 160 000 ms 180 Status Initiato Time 200 B 442 me

(index):27

websock... react refresh:6

~

G 🖈 🖸 🌖 :

302 B 10 ms

0B Pending

363 B 21 ms

204			100000		
304	manifest	Other	363 B	23 ms	
304	png	Other	364 B	3 ms	
403	media	Other	213 B	123 ms	
200	ipeg	react-dom-client.de	08	0 ms	
200	fetch	App.is:32	252 B	47.27 s	
403	media	Other	213 B	124 ms	
200	ipeg	react-dom-client.de	08	2 ms	
403	media	Other	213 B	114 ms	
200	ipeg	react-dom-client.de	0.8	0 ms	
200	fetch	Ano.is.32	249 B	35.25 s	
	304 403 200 200 403 200 403 200 200	Jua mantest 403 media 403 media 200 jeg 200 fetch 403 media 200 jeg 200 jeg 200 fetch	alua manines Onint Maria Dinter media Onint media Onint Resolution i pege residuation de la consultant de Resolution de la consultant Resolution de la consultant Resoluta	au materials Coher 483 B 403 media Coher 234 B 403 media Coher 213 B 200 lipeig asci_star_star_star_star_star_star_star_star	Aud mediens Onter 349 8 A ann 403 media Other 2138 123 me 403 media Other 2138 123 me 300 large stackdom:jettde 08 0 me 300 media Other 2138 1218 124 me 300 media Other 2138 1218 1218 1218 1218 1218 1218 1218

304

14 requests 2.8 k8 transferred 1.4 MB resources Finish: 2.8 min DOMContentLoaded: 2.34 s Lt

 $\leftarrow \rightarrow C$ (D) localhost:3000

🗧 💿 🌒 🎲 Topic: Midtern 🗴 🕲 CS252-Day13- 🗴 🥥 CS252-Day13- X 😻 CS252-Day13- X 🖤 Vite + React X 📓 React App 🙆 X 🍵 post-completi 🛛 X 👩 Post-Completi 🖉 X

Capture Image Close Camer

* Preview:



Taption: a young girl is making a funny face while standing in a room.

K LD Elements Console St	urces N	etwork Pe	rformance >>	0 16 8	3 : X
🖲 🖉 🔻 🤉 🗆 Preserve log	🗌 Disat	ole cache N	o throttling 🔹 🗟	: 1 ±	۲
Y Filter	nvert M	ore filters +			
All Fetch/XHR Doc CSS JS F	ont Img	Media Mar	ifest WS Wasm	Other	
200,000 ms 400,000 ms	600,000 ms	800,00	0 ms 1,000,000 mi	1,200,0	00 ms
		Ten			
		-	E	-	
Name	Status	Туре	Initiator	Size	Time
bundle.is	304	script	(index):27	302 B	10 m
+* ws	101	websock	react refresh:6	0.8	Pendir
favicon.ico	304	x-icon	Other	363 B	21 n
🗅 manifest.json	304	manifest	Other	363 B	23 n
logo192.png	304	png	Other	364 B	3 n
8 beep-07.wav	403	media	Other	213 B	123 n
blob:http://localhost:3000/94e63	200	ipeg	react-dom-client.de	0 B	0 m
> caption/	200	fetch	App.is:32	252 B	47.27
beep-07.wav	403	media	Other	213 B	124 r
blob:http://localhost:3000/f384cb	200	jpeg	react-dom-client.de	08	21
beep-07.wav	403	media	Other	213 B	114 n
blob:http://localhost:3000/b8e52	200	jpeg	react-dom-client.de	08	0 m
(i) caption/	200	fetch	App.is:32	249 B	35.25
beep-07.wav	403	media	Other	213 B	114 r
blob:http://localhost:3000/ac448f	200	jpeg	react-dom-client.de	0 B	1 m
() caption/	200	fetch	App.is:32	252 B	25.02
beep-07.wav	403	media	Other	213 B	110 m
blob:http://localhost:3000/8c91b9	200	jpeg	react-dom-client.de	0 B	2 n
02 caption/	200	fetch	App.is:32	261 B	21.57
beep-07.wav	403	media	Other	213 B	111 n
blob:http://localhost:3000/ee873a	200	jpeg	react-dom-client.de	0 B	2 n
O caption/	200	fetch	App.is:32	257 B	25.69
S beep-07.wav	403	media	Other	213 B	139 n
# blob:http://localhost:3000/b75d5	200	jpeg	react-dom-client.de	0 B	2 n
beep-07.wav	403	media	Other	213 B	153 n
blob:http://localhost:3000/3fdaf6	200	jpeg	react-dom-client.de	08	1 n
(i) caption/	200	fetch	App.js:32	272 B	12.31

C + D 6 :

GIT-base Architecture and Processing Flow – How it works

Vision Encoder: ViT (Vision Transformer)

- Splits the image into fixed-size patches (e.g., 16×16)
- Adds position embeddings
- Passes through multi-head self-attention layers
- Outputs a sequence of visual tokens
- Text Decoder: Transformer-based decoder
- Receives visual tokens as context (via cross-attention)
- Generates text token-by-token using masked self-attention
- Output is autoregressive (each word depends on previous output)

GIT-base Architecture and Processing Flow – How it works



GIT-base Processing Pipeline

- Resize the input image to 224×224 (or as required by model config)
- Normalize pixel values to match pretrained weights
- Divide into patches (e.g., 14×14 = 196 total patches for 16×16 patch size)
- Flatten and project patches to a fixed embedding dimension (e.g., 768)
- Add positional encodings
- Pass into ViT encoder
- Feed encoded visual sequence into decoder
- Caption generation starts with a start token <bos>
- Generates next tokens until end-of-sequence <eos>

GIT-base Processing Pipeline

Training Details

- Pretrained on:
 - COCO dataset: 120k+ images with 5 human-written captions each
 - Conceptual Captions: 3.3M noisy web images with alt-text captions
- Uses cross-entropy loss during training for next-token prediction

Frontend- App.js

Technologies Used

- **React.js** (functional components + hooks)
- HTML5 Canvas API for capturing webcam frames
- Web Speech API for text-to-speech output
- JavaScript Events for keyboard and mobile tap interaction









Key Functionalities in App.js

Function	What It Does
handleCapture()	Captures image from webcam via <canvas></canvas>
sendToBackend()	Sends image as blob to FastAPI /caption
speak(caption)	Uses Web Speech API to speak caption
useEffect()	Detects mobile or desktop and sets up camera
handleTap()	Detects mobile tap for image capture
playSound(type)	Plays feedback sounds: capture, loading, success

Accessibility Features

- No instructions needed app speaks to user on its own:
 - "Tap anywhere to open the camera" (on mobile)
 - "Tap anywhere to generate caption" (on Desktop)
- Voice Feedback via Web Speech API:
 - Speaks welcome message
 - Speaks when caption is being generated
 - Speaks the final caption
 - Speaks error messages if API fails
- Auditory Feedback:
 - **Beep** = Image captured
 - Ticking = Processing
 - **Ding** = Caption ready



Accessibility Features

- Camera Auto-Config:
 - Uses navigator.userAgent to detect mobile/desktop
 - Adjusts preview size accordingly
- Visual Feedback:
 - Blue glowing outline around live video to indicate camera is active
 - Button flash or highlight when tapping screen
- Mobile-First Support:
 - Tap anywhere to take photo
 - Single-screen fullscreen mode
 - No buttons needed for users with low vision



Backend – app.py

Backend Technology

- Framework: FastAPI (Python)
- Role: Acts as a lightweight bridge between frontend and Hugging Face
- Model Hosting: Offloaded to Hugging Face Inference API (microsoft/git-base-coco) Processing Flow
- 1. Frontend sends captured image via a POST request to /caption/
- 2. FastAPI backend receives image in binary (bytes) format
- 3. Backend forwards image to Hugging Face model API with authorization token
- 4. Receives caption output as JSON from Hugging Face
- 5. Sends **parsed caption** back to frontend for voice + display





Backend – app.py

Security + Reliability

- Uses HF_TOKEN stored as environment variable (not hardcoded)
- Enables CORS middleware to allow cross-origin requests from frontend
- Error handling:
 - Catches failed requests
 - Sends clear error messages (e.g., "Image processing failed")
 - Tries to parse only valid JSON response
 - Offloads heavy ML inference to Hugging Face cloud



Deployment – Frontend on Vercel

- React frontend deployed on Vercel
- Project pushed to **GitHub** → Linked with Vercel for auto-deployment
- Updated BACKEND_URL in frontend to point to live Render backend
- Ran npm run build for optimized production bundle
- Vercel auto-detected framework and used global **CDN** to serve app fast
- Publicly accessible at https://visionmate-theta.vercel.app/
- No login/authentication required
- Works on desktop and mobile (including slow networks)
- Instant response within a few seconds per caption

Deployment – Frontend on Vercel



Deployment – Backend on Render

Backend built with FastAPI (app.py)

Initially tried model loading locally, but Render's 512MB RAM limit caused:

- Crashes
- Missing CPU instruction errors

Switched to **API-based approach** using Hugging Face's hosted model (git-base-coco) Backend acts as a relay: receives image \rightarrow forwards to Hugging Face \rightarrow returns caption Set HF_TOKEN as **environment variable** (not in code) Enabled **CORS** for frontend communication

• Live API endpoint: https://visionmate-backend.onrender.com/caption/

Deployment – Backend on Render

< > C	5 https://dashboard.render.com/web/srv-cvso363uibrs73ectvf0 🖞 🚺 🚺 🗘 🚺 🖬 😓 🕡 Update 🚍
🚫 🛛 🧵 Academic History 😭 Meet - yfg	g-ckzw-u 🧃 Meet - ueu-ozrk-cor 🕺 Untitled - Jupyter 🐨 Python Basic: Exer 💟 Manvitha's Nithin J 💶 (321) Krishna Cart 🔅 🗎 🕮 All Bookmarks
My Workspace 🗘	visionmate-backend V Q Search KK + New G Upgrade O S
← Dashboard ⊕ visionmate-backend	web service visionmate-backend Connect ~ Manual Deploy ~
i≡ Events	Docker Free Upgrade your instance →
Settings	🗘 sai-anoushka / visionmate 🗠 🗠 master
MONITOR	https://visionmate-backend.onrender.com
™o Logs	
⊭ Metrics	(a) Your free instance will spin down with inactivity, which can delay requests by 50 seconds or more.
MANAGE	Deploy live for <u>offb267</u> ; Final push hopefullygit add . April 22, 2025 at 118 AM
♂ Scaling ♥	Deploy started for <u>cffb267</u> . Final push hopefullygit add .
Previews	leÎ Manually triggered by you via Dashboard
🗧 Disks 🔸	April 22, 2025 at 115 AM
🗅 Jobs 🕈	Deploy live for <u>9051455</u> : Initial clean version of VisionMate
E Changelog	April 20, 2025 at 6:05 PM
 ✓ Invite a friend Contact support 	Deploy started for <u>9951455</u> : Initial clean version of VisionMate © New commit via Auto-Deploy
√ Render Status	אַטָּדוו געו, בעבל פֿג טעש דיאו

Desktop View

UisionMate - Image Captioning Choose file No file chosen Generate Caption Open Camera (Laptop)	VisionMate - Image Captioning			19		OFV 22 25	. : X
VisionMate - Image Captioning Choose file No file chosen Copen Camera (Laptop) Name Status	VisionMate - Image Captioning	I I I I I PIENEL	ve log Disa	ble cache N	lo throttling		R
Chooses file No file chosen Generate Caption Open Camera (Laptop) 90 ms 100 ms 100 ms 20 ms 20 ms 20 ms Name Status Type Initiator Size Time Cocahost 304 document 01ms 200 ms	rioloninato intago captioning			lore filters -			~
Choose file No file chosen Cenerate Caption Open Camera (Laptop) 00 ms 100 ms 100 ms 200 ms <t< td=""><td></td><td>All Fatably U.D. Day (CCC)</td><td>Cant lan</td><td>Madia Mar</td><td>ifant WC Man</td><td>Other</td><td></td></t<>		All Fatably U.D. Day (CCC)	Cant lan	Madia Mar	ifant WC Man	Other	
Open Camera (Laptop) Name Status Type Initiator Stare Time It caclhoatt 304 societ Other 290 34 It caclhoatt 304 societ Other 290 34 It caclhoatt 304 societ Other 3028 7 mm It societ 304 societ Other 363.8 8 mm It andreation 304 manifest Other 363.8 5 mm It age/cpi/2png 304 png Other 364.8 5 mm	Choose file No file chosen Generate Caption	50 ms	100 ms	150 ms	200 ms	Other	50 ma
NameStatusTypeInitiatorSizeTimeIS localhost304documentOther29934 mIS boundle js304scriptIfindeni.323027 mI' wo101webcck.most.frattenist.304 %script363 %8 mI' avison.loco304x-loonOther363 89 m363 %8 mI' marifest.jon304pngOther363 89 mI' logo192.png304pngOther364 %5 m	Open Camera (Laptop)						
□ cacheat 394 obcument 01er 299 B 34 m □ bundle js 304 sich findsab22 302 B 7 m =' ws 101 websock_model_taffsab32 302 B 7 m =' ws 101 websock_model_taffsab32 303 B 8 m □ favionico 304 x-icon Other 363 B 9 m □ favionico 304 marfest Other 363 B 9 m □ favionico 304 png Other 364 B 5 m □ favionico 104 png 364 B 5 m		Name	Status	Туре	Initiator	Size	Time
□ pundie,js 30,4 seript findex1,22 30,28 > 7 manifest ■ twoch.co 101 webck. treat/rest.frefst.36 0.8 Pendin ■ twoch.co 30.4 s.4-con Other 335.8 8 manifest □ manifest.juon 30.4 pmg Other 335.8 9 manifest ■ logo192.prg 30.4 pmg Other 364.8 5 manifest		Iocalhost	304	document	Other	299 B	34 m
i* vs 101 websockreak_traftesb.5 008 Pendin If aviconico 304 :-con Other 335 8 m In manifest juon 304 manifest Other 335 9 m is logo192.png 304 png Other 3364 5 m		🗈 bundle.js	304	script	(index):27	302 B	7 m:
∎ fravisonico 304 x-icon Other 363 8 8m ⊡ marifest.jon 304 png Other 363 9m ■ legot92.png 304 png Other 364 8 5m		≓ ws	101	websock	react refresh:6	0 B	Pending
D manfest.jon 304 manfest Other 383 8 9m ■ logo192.png 304 png Other 3648 5m		favicon.ico	304	x-icon	Other	363 B	8 m
logo192.png 304 png Other 364.8 5m		🗅 manifest.json	304	manifest	Other	363 B	9 m

6 requests 1.7 kB transferred 1.3 MB resources Finish: 197 ms DOMContentLoaded: 134 ms

Desktop View

🖲 💿 😳 SJSI X 🥥 CS2 X 🕲 Chri X 🕲 Moci X 🕲 cs.s; X 🕲 latti: X 📓 F 🖲 X 🔿 C	SitH 🗙 🎒 Spri 🗙 🕘 cs.s 🗙	🖸 Lect 🗙 💙	Vite × 🔇 127.	× +						
← → C 💿 localhost:3000			C \$	5	S :					
	Elements Console	ources Network	Performance Mem	ory >> 🔅	: ×					
		Disable cache	No throttling	6 1 4	æ					
VisionMate - Image Captioning										
	Y Filter									
Choose file No file chosen Generate Caption	All Fetch/XHR Doc CSS JS	Font Img Media	Manifest WS Wasm	Other						
	50,000 ms 100,000 ms	150,000 ms 200,0	0 ms 250,000 ms	300,000 ms	350,0					
	Name	Status Type	Initiator	Size	Time					
	Iocalhost	304 docum	nt Other	299 B	332 ms					
	🙂 bundle.js	200 script	(index):27	239 kB	63 ms					
	∓* ws	101 webso	k react refresh:6	0 B	Pending					
	a favicon.ico	304 x-icon	Other	363 B	8 ms					
	D manifest.json	304 manife	t Other	363 B	10 ms					
	logo192.png	304 png	Other	364 B	4 ms					
	blob:http://localhost:3000/05ccdb.	200 jpeg	react-dom-client.c	de OB	0 ms					
	() caption/	200 fetch	App.is:35	255 B	2.79 s					
	blob:http://localhost:3000/f1f9f48.	. 200 jpeg	react-dom-client.c	<u>i</u> € 0 B	4 ms					
	caption/	200 fetch	App.js:35	251 B	2.90 s					
Capture image Close Camera	blob:http://localhost:3000/58eed	200 jpeg	react-dom-client.c	<u>ie</u> 0 B	1 ms					
Providence	(i) caption/	200 fetch	App.is:35	249 B	908 ms					
Preview:	blob:http://localhost:3000/3ecc19.	. 200 jpeg	react-dom-client.c	<u>ie</u> 0 B	1 ms					
	(i) caption/	200 fetch	App.js:35	251 B	689 ms					
	blob:http://localhost:3000/76b6b	200 jpeg	react-dom-client.c	15 O B	0 ms					
	() caption/	200 fetch	App.is:35	251 B	692 ms					
	blob:http://localhost:3000/799cb	200 jpeg	react-dom-client.c	1 <u>€</u> 0 B	1 ms					
The second	 caption/ 	200 fetch	App.is:35	249 B	706 ms					
	()) caption/	200 fetch	App.is:35	249 B	787 ms					
	hlob:http://localhost:3000/25388	200 ipeg	react-dom-client of	le OB	0 ms					

() caption

()) caption/

() caption

blob:http://localhost:3000/8d366...

blob:http://localhost:3000/23903.

Caption: a woman wearing sunglasses and holding a cell

25 requests 243 kB transferred 1.5 MB resources Finish: 5.4 min DOMContentLoaded: 478 ms

fetch

inea

200 ipeg

200

200 fetch

200

App.is:35

App. js:35

App.is:35

react-dom-client.de

react-dom-client

930 ms

1.56 s

4 ms

252 B

251 B

O F

256 B 2.89 s

0 B 11 ms

Desktop View



Caption:

Desktop View



Desktop View



Desktop View



Mobile View





Caption: a girl with a laptop



Mobile View



Mobile View



Caption: a photo of a laptop screen with a woman on it.

10.0.0.69

Mobile View



Mobile View





Testing and Performance Evaluation

What was tested?

- Camera capture response time
- Caption generation latency
- Speech output delay
- **Device compatibility** (Desktop + Mobile)
- Browser support: Chrome, Safari, Firefox
- Model performance: BLIP-base vs GIT-base

Testing and Performance Evaluation

Test Type	Result
Average caption generation time	5–7 seconds (GIT-base via HF API)
Audio output delay	~0.5 seconds after caption response
Success rate (20 custom images)	18/20 images received relevant captions
Browser support	Chrome, Firefox, Safari (Mobile/Desktop)
Caption accuracy (BLIP vs GIT)	GIT-base was more consistent for real scenes

Limitations

Latency in First Request:

- Render's free backend spins down after inactivity.
- First API call may take **20–40 seconds** to respond.

Hugging Face API Rate Limits:

- Can break under heavy usage.
- Hugging Face has the ability to remove API inference anytime they please. Very Dependent

Internet Dependency:

- Requires stable internet on both frontend and backend for full functionality.
- Not usable offline due to reliance on external model API.

Caption Ambiguity:

• GIT-base sometimes generates vague or overly broad descriptions (e.g., "a person in a room").

No Real-Time Video Support Yet:

- App only works on single images, not continuous frames or live streaming. **No Multilingual Support**:
- Currently limited to English text and speech output.

Conclusion

- VisionMate enables image-based understanding using AI-generated captions, especially designed for users with visual impairments.
- Combines a **React.js** frontend (camera, TTS, mobile/desktop support) with a **FastAPI** backend connected to **Hugging Face's GIT-base model**.
- Deployed using **Vercel** and **Render**, both on free plans.
- GIT-base selected for speed (5–8s avg.) and accuracy in real-world testing.

Future work

- Video Captioning Support:
 - Modern videos are 30–50 frames per second (fps).
 - VisionMate could be extended to capture keyframes from short videos (e.g., 1 frame every second) and generate a caption per scene.
 - Useful for summarizing scenes or assisting users in real-time visual navigation (e.g., short video clips, walking around a room).
- Custom Model Fine-Tuning:
 - Train GIT or BLIP model on specific domains (e.g., indoor, educational settings)
- Multilingual Captioning:
 - Add translations and voice output in multiple languages.

Future work

- Voice-Controlled Interaction:
 - Let users say "capture", "read", or "exit" instead of tapping/clicking
- Better Accessibility Integration:
 - Add semantic roles, and support for screen readers like VoiceOver and TalkBack
- Offline Functionality:
 - Use small models or preloaded ML to run on-device without Internet

References

1] J. Li, et al., "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," arXiv preprint arXiv:2201.12086, 2022. Available: https://arxiv.org/abs/2201.12086

[2] L. H. Wang, et al., "GIT: A Generative Image-to-Text Transformer for Vision-to-Language Pretraining," arXiv preprint arXiv:2205.14100, 2022. Available: <u>https://arxiv.org/abs/2205.14100</u>
[3] T. Y. Lin, et al., "Microsoft COCO: Common Objects in Context," ECCV, 2014. Available: https://cocodataset.org/

[4] S. Ramírez, "FastAPI Documentation," FastAPI, [Online]. Available: https://fastapi.tiangolo.com/

[5] T. Wolf, et al., "Transformers: State-of-the-Art Natural Language Processing," EMNLP, 2020. Available: <u>https://huggingface.co/docs/transformers/index</u>

[6] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no.8, pp. 1735–1780, 1997. Available: https://www.bioinf.jku.at/publications/older/2604.pdf

Thank you !

Questions?