GPT-4 Turbo with Vision concepts

02/08/2024

GPT-4 Turbo with Vision is a large multimodal model (LMM) developed by OpenAI that can analyze images and provide textual responses to questions about them. It incorporates both natural language processing and visual understanding. This guide provides details on the capabilities and limitations of GPT-4 Turbo with Vision.

To try out GPT-4 Turbo with Vision, see the quickstart.

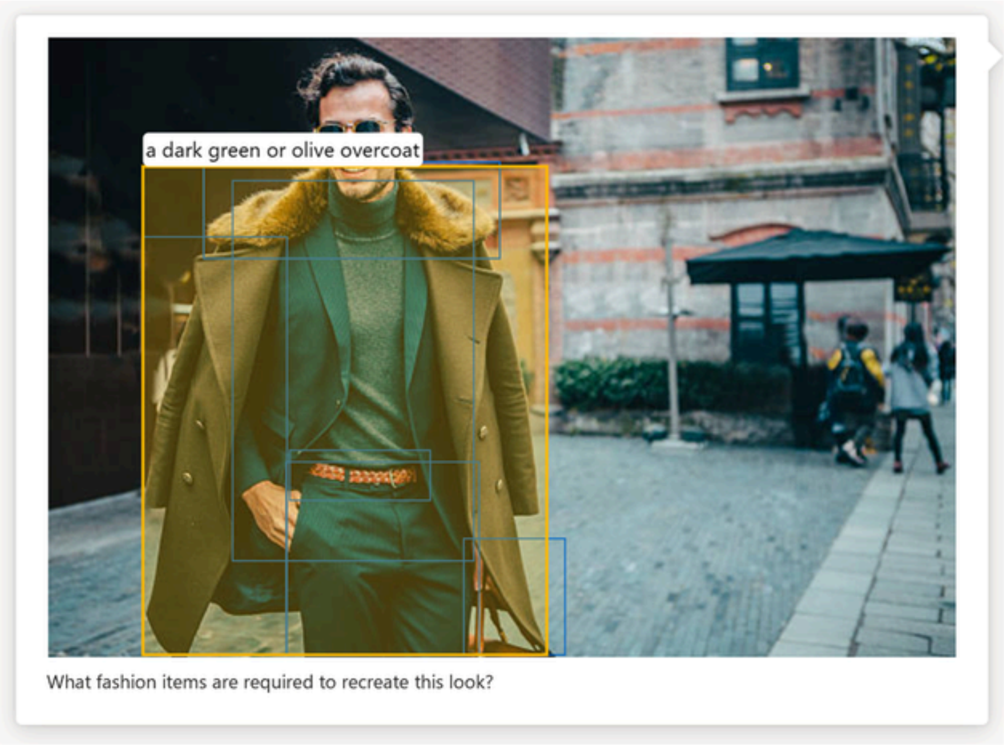
The GPT-4 Turbo with Vision model answers general questions about what's present in the images or videos you upload.

Enhancements let you incorporate other Azure AI services (such as Azure AI Vision) to add new functionality to the chat-with-vision experience.

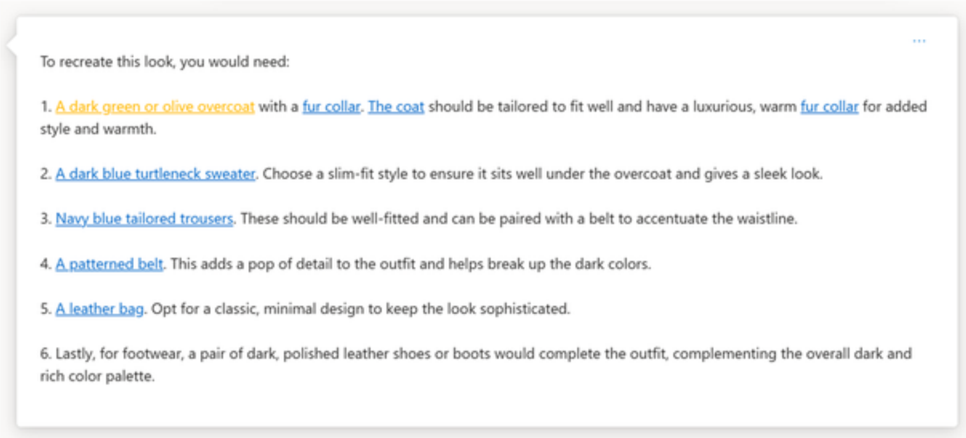
Object grounding: Azure AI Vision complements GPT-4 Turbo with Vision's text response by identifying and locating salient objects in the input images. This lets the chat model give more accurate and detailed responses about the contents of the image.

Important

To use Vision enhancement, you need a Computer Vision resource. It must be in the paid (S1) tier and in the same Azure region as your GPT-4 Turbo with Vision resource.



What fashion items are required to recreate this look?



To recreate this look, you would need:

1. A dark green or olive overcoat with a fur collar. The coat should be tailored to fit well and have a luxurious, warm fur collar for added style and warmth.

2. A dark blue turtleneck sweater. Choose a slim-fit style to ensure it sits well under the overcoat and gives a sleek look.

3. Navy blue tailored trousers. These should be well-fitted and can be paired with a belt to accentuate the waistline.

4. A patterned belt. This adds a pop of detail to the outfit and helps break up the dark colors.

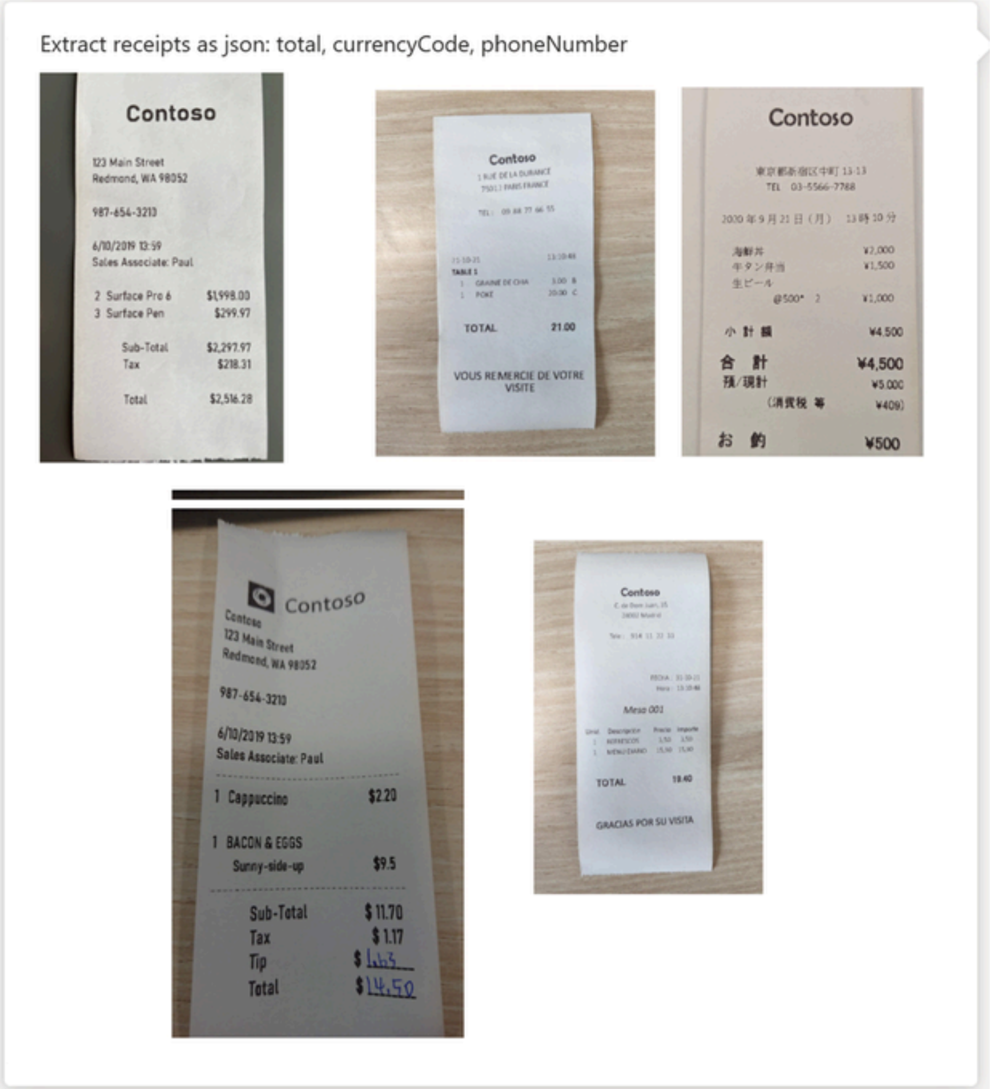
5. A leather bag. Opt for a classic, minimal design to keep the look sophisticated.

6. Lastly, for footwear, a pair of dark, polished leather shoes or boots would complete the outfit, complementing the overall dark and rich color palette.

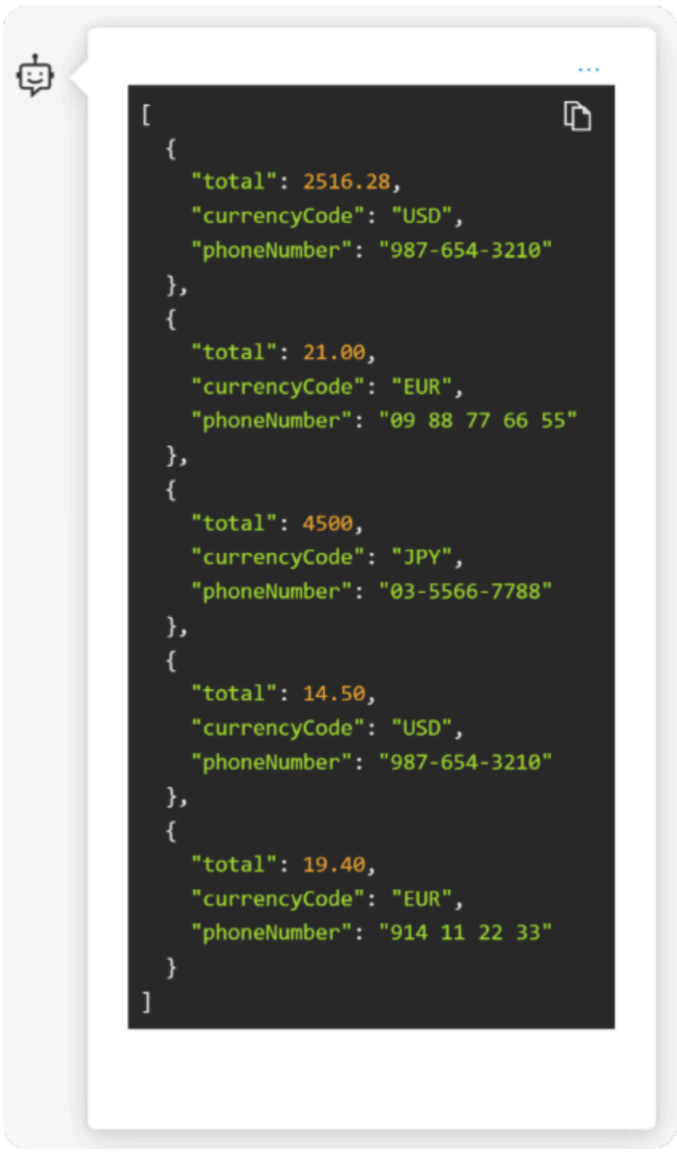
Optical Character Recognition (OCR): Azure AI Vision complements GPT-4 Turbo with Vision by providing high-quality OCR results as supplementary information to the chat model. It allows the model to produce higher quality responses for images with dense text, transformed images, and numbers-heavy financial documents, and increases the variety of languages the model can recognize in text.

Important

To use Vision enhancement, you need a Computer Vision resource. It must be in the paid (S0) tier and in the same Azure region as your GPT-4 Turbo with Vision resource.



Extract receipts as json: total, currencyCode, phoneNumber



```
[
  {
    "total": 2516.28,
    "currencyCode": "USD",
    "phoneNumber": "987-654-3210"
  },
  {
    "total": 21.00,
    "currencyCode": "EUR",
    "phoneNumber": "09 88 77 66 55"
  },
  {
    "total": 4500,
    "currencyCode": "JPY",
    "phoneNumber": "03-5566-7788"
  },
  {
    "total": 14.50,
    "currencyCode": "USD",
    "phoneNumber": "987-654-3210"
  },
  {
    "total": 19.40,
    "currencyCode": "EUR",
    "phoneNumber": "914 11 22 33"
  }
]
```

Video prompt: The **video prompt** enhancement lets you use video clips as input for AI chat, enabling the model to generate summaries and answers about video content. It uses Azure AI Vision Video Retrieval to sample a set of frames from a video and create a transcript of the speech in the video.

Note

In order to use the video prompt enhancement, you need both an Azure AI Vision resource and an Azure Video Indexer resource, in the paid (S0) tier, in addition to your Azure OpenAI resource.

Special pricing information

Important

Pricing details are subject to change in the future.

GPT-4 Turbo with Vision accrues charges like other Azure OpenAI chat models. You pay a per-token rate for the prompts and completions, detailed on the Pricing page. The base charges and additional features are outlined here:

Base Pricing for GPT-4 Turbo with Vision is:

- Input: $0.01 per 1000 tokens
- Output: $0.03 per 1000 tokens

See the Tokens section of the overview for information on how text and images translate to tokens.

If you turn on Enhancements, additional usage applies for using GPT-4 Turbo with Vision with Azure AI Vision functionality.

Model	Price
+ Enhanced add-on features for OCR	$1.5 per 1000 transactions
+ Enhanced add-on features for Object Detection	$1.5 per 1000 transactions
+ Enhanced add-on feature for “Add your Image” Image Embeddings	$1.5 per 1000 transactions
+ Enhanced add-on feature for “Video Retrieval” integration [1]	Ingestion: $0.05 per minute of video Transactions: $0.25 per 1000 queries of the Video Retrieval index

[1] Processing videos involves the use of extra tokens to identify key frames for analysis. The number of these additional tokens will be roughly equivalent to the sum of the tokens in the text input, plus 700 tokens.

Example image price calculation

Important

The following content is an example only, and prices are subject to change in the future.

For a typical use case, take an image with both visible objects and text and a 100-token prompt input. When the service processes the prompt, it generates 100 tokens of output. In the image, both text and objects can be detected. The price of this transaction would be:

Item	Detail	Total Cost
GPT-4 Turbo with Vision input tokens	100 text tokens	$0.001
Enhanced add-on features for OCR	$1.50 / 1000 transactions	$0.0015
Enhanced add-on features for Object Grounding	$1.50 / 1000 transactions	$0.0015
Output Tokens	100 tokens (assumed)	$0.003
Total Cost		$0.007

Example video price calculation

Important

The following content is an example only, and prices are subject to change in the future.

For a typical use case, take a 3-minute video with a 100-token prompt input. The video has a transcript that's 100 tokens long, and when the service processes the prompt, it generates 100 tokens of output. The pricing for this transaction would be:

Item	Detail	Total Cost
GPT-4 Turbo with Vision input tokens	100 text tokens	$0.001
Additional Cost to identify frames	100 input tokens + 700 tokens + 1 Video Retrieval transaction	$0.00825
Image Inputs and Transcript Input	20 images (85 tokens each) + 100 transcript tokens	$0.018
Output Tokens	100 tokens (assumed)	$0.003
Total Cost		**$0.03025**

Additionally, there's a one-time indexing cost of $0.15 to generate the Video Retrieval index for this 3-minute video. This index can be reused across any number of Video Retrieval and GPT-4 Turbo with Vision API calls.

This section describes the limitations of GPT-4 Turbo with Vision.

- **Limitation on image enhancements per chat session**: Enhancements cannot be applied to multiple images within a single chat call.
- **Maximum input image size**: The maximum size for input images is restricted to 20 MB.
- **Object grounding in enhancement API**: When the enhancement API is used for object grounding, and the model detects duplicates of an object, it will generate one bounding box and label for all the duplicates instead of separate ones for each.
- **Low resolution accuracy**: When images are analyzed using the "low resolution" setting, it allows for faster responses and uses fewer input tokens for certain use cases. However, this could impact the accuracy of object and text recognition within the image.

- **Image chat restriction**: When you upload images in Azure OpenAI Studio or the API, there is a limit of 10 images per chat call.

- **Low resolution**: Video frames are analyzed using GPT-4 Turbo with Vision's "low resolution" setting, which may affect the accuracy of small object and text recognition in the video.
- **Video file limits**: Both MP4 and MOV file types are supported. In Azure OpenAI Studio, videos must be less than 3 minutes long. When you use the API there is no such limitation.
- **Prompt limits**: Video prompts only contain one video and no images. In Azure OpenAI Studio, you can clear the session to try another video or images.
- **Limited frame selection**: The service selects 20 frames from the entire video, which might not capture all the critical moments or details. Frame selection can be approximately evenly spread through the video or focused by a specific video retrieval query, depending on the prompt.
- **Language support**: The service primarily supports English for grounding with transcripts. Transcripts don't provide accurate information on lyrics in songs.

- Get started using GPT-4 Turbo with Vision by following the quickstart.
- For a more in-depth look at the APIs, and to use video prompts in chat, follow the how-to guide.
- See the completions and embeddings API reference