# Understand multimodal prompting

If you were an artist, you could probably create a masterpiece using only one color of paint—but using more colors gives you way more power to produce something amazing. The same is true when it comes to a prompt. You can accomplish a lot with text-only prompts, but you can add more elements to take them to the next level. With multimodal prompting, you can achieve much more than with text alone, opening up a world of new possibilities.



## **What is multimodal prompting?**

Multimodal prompting is a big term with a simple meaning. To break it down, modalities are different types of formats, like text, image, or audio, that can be used to instruct a generative AI tool. So multimodal prompting is when you use multiple modalities in the same prompt.

Multimodal prompting is useful because it lets you do things like translate images into detailed text descriptions for visually impaired users, or generate graphics based on an audio recording that describes a marketing campaign.

## **Benefits and considerations**

Think about it like this: the world is multimodal. When you’re giving a presentation at work, you’re probably using text, images, and your own voice to help your audience understand your point. Multimodal prompting works in the same way. It reflects how you experience the world by building connections between text, images, audio, and other modalities, helping you provide clearer instructions to a gen AI tool.

Including multiple modalities isn’t a one-size-fits-all solution, but it can improve your prompts. For instance, if you ask a gen AI tool to write a poem about a sunset and include a photo of one in your input, the gen AI tool can use the photo to describe specific colors and details. The result is a much more vivid poem.

While multimodal prompting is versatile, it has some limitations. Gen AI tools sometimes struggle with abstract ideas, handling complex combinations of modalities, and occasionally, just plain old common sense. For example, while gen AI tools are always evolving and improving, multimodal prompting isn’t reliable for counting just yet.

## **Multimodal prompting examples**

Multimodal prompting is most effective in situations where you need to transfer information from one media to another. Here are a few examples:

* You ask a gen AI tool to suggest recipes based on a photo of the ingredients in your fridge.
* You need to create a digital teaser to promote an event between two brands on social media. So you input both brands’ logos and colors into a gen AI tool, and ask it to generate a visual for the joint event.
* You’re working on a short story about a forest you just visited and need help describing the sounds and atmosphere. So you add some audio you recorded while visiting the forest to your prompt, and ask the gen AI tool to use it to inspire the atmosphere and details of the story.

## **Using the prompting framework**

Just like when you design text-based prompts, you’ll still rely on the **prompting framework** for multimodal prompts: **T**houghtfully **C**reate **R**eally **E**xcellent **I**nputs. However, you might need to adjust the framework based on the modalities you’re using in your prompt.

* ***Task***: Along with being specific about a ***persona*** and ***format***, be specific about how you want different modalities to be used and why you’re including them in the task.
* ***Context***: Just like in a text-based prompt, you’ll need to add context so the gen AI tool knows exactly what you want it to do. Consider which modalities you’re using in your prompt when providing context and if there are any specific constraints or considerations that are important to include. For example, if you input an image of the inside of your refrigerator, you may only want the gen AI tool to consider the fruits and vegetables in the image and not any of the packaged goods.
* ***References***: Share written, visual, or audio references to help gen AI tools understand the type of output you’re trying to create. Not every tool supports every sort of reference, so make sure you’re using a gen AI tool that supports the references you want to give it.
* ***Evaluate***: Run the prompt as usual, but take extra care to evaluate how the output looks or reads. Since you’re using different modalities, that might involve a more detailed review of the output than simply evaluating a text response.
* ***Iterate***: If the output is not meeting your needs, try one of the four iteration methods: you can revisit the prompting framework, break the prompt down into shorter tasks, introduce constraints, or tweak your phrasing or switch to an analogous task to help trigger a different, potentially better, response.

Multimodal prompting is a powerful tool that lets you transfer information into different mediums and create endless possibilities of outputs. You’re the artist, and the gen AI tool is your paintbrush.