Edge AI Sizing Tool

Get Started Guide

June 2025

Document Number: 814160-2.0

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted, which includes subject matter disclosed herein.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or visiting the Intel Resource and Documentation Center.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product or component can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com .

No product or component can be absolutely secure.

Contents

Figures

Tables

Revision History

Date	Revision	Description
May 2025	2.0	Open-Source Release.
October 2024	1.1	Evaluation release.
February 2024	1.0	Non-production initial release.

§

1.0 Introduction

The Edge AI Sizing Tool has been designed to assist users in sizing AI models for edge devices, which often have limited computational resources such as processing power, memory, and storage. This tool provides a zero-code configuration and an easy-to-use interface that permits users to effortlessly set up AI applications by selecting inputs, accelerators, performance modes, and AI models. It also offers real-time monitoring of system performance metrics, such as CPU and GPU usage, memory consumption, and inference speed, which enables users to optimize AI workflows and make informed decisions. This tool enhances performance and user experience by ensuring that AI models are scaled efficiently to operate within constraints. Finally, this release has a fresh and sleek design look compared to its previous iterations, where the UI is sleeker and more user-friendly.

For release information and notes, refer to Release Note 4.0 Release Notes.

1.1 Terminology

Table 1. Terminology

Abbreviation	Description
AI	Artificial Intelligence
GUI	Graphical User Interface
OpenVINO™	Opensource toolkit used for optimizing and deploying AI interface
LLM	Large Language Models
OS	Operating System
UI	User Interface
GB	Gigabytes
RAM	Random Access Memory
CPU	Central Processing Unit
iGPU	Integrated Graphics Processing Unit
dGPU	Discrete Graphics Processing Unit
NPU	Neural Processing Unit
NLP	Natural Language Processing

Abbreviation	Description
FPS	Frames Per Second

1.2 System Requirements

You will require a system that meets the following requirements:

- **Processor:**
  - Intel® Core™ Ultra Processors (Series 2) (Minimum Requirement)
  - 12th Generation Intel® Core™ Processors or above (Recommended)
- **Internet Connection:** Required for downloading models
- **OS:**
  - Ubuntu 24.04 LTS Desktop (Recommended)
  - Ubuntu 22.04 LTS Desktop
- **Disk Space:** 256 GB free disk size (Minimum Requirement)
- **RAM:**
  - 8 GB (Minimum Requirement)
  - 32 GB or higher (Recommended)
- **Optional Devices:**
  - Intel® Graphics Compute Runtime for oneAPI Level Zero and OpenCL (TM) Driver 25.09.32961.5
  - Intel NPU driver v1.16.0

1.3 Validated Configurations

Edge AI Sizing Tool has been validated on the following:

Table 2. Validated Configuration Details

Hardware	• CPU: Intel® Core™ Ultra 7 165H (Series 1)
	• GPU:
	  ○ Intel® Arc™ A770 Graphics (dGPU)
	  Intel® Arc™ Graphics (iGPU)
	• NPU: Intel® AI Boost
	• RAM: 64GB
	• Disk Capacity: 256GB

Software	• OSs: Ubuntu* 24.04 LTS Desktop (Linux)
	• Package:
	  ○ Python: 3.12
	  ○ Node.js: 22.15.1
	  ○ OpenVINO™ toolkit 2025.0.0
	  ○ OpenVINO™ toolkit GenAI 2025.1.0
	  ○ Optimum Intel 1.22.0
	  ○ Ultralytics 8.3.61

§

2.0 Pre-requisites

This section covers the necessary preparation for setting up the Edge AI Sizing Tool.

2.1 Operating System

The Edge AI Sizing Tool is compatible with the OSes below:

- Ubuntu* 24.04 & Ubuntu* 22.04 (Refer to https://ubuntu.com/)

2.2 Proxy Configuration

Note: Skip this point if your system is not behind a proxy network.

Ensure that the proxy addresses are configured in the system. An example of configuring the proxy network is shown below:

1. Open the environment *(/etc/environment)* file, use a text editor or a terminal with root user privilege to open it:

```
$ sudo nano /etc/environment
```

2. Add the proxy variables as shown in the figure below and modify them to fit your environment.

Figure 1. Example of Proxy Settings for Ubuntu* OS

```
http_proxy=http://<username>:<password>@<hostname>:<port>
https_proxy=http://<username>:<password>@<hostname>:<port>
no_proxy=localhost,<pattern>,...
HTTP_PROXY =http://<username>:<password>@<hostname>:<port>
HTTPS_PROXY =http://<username>:<password>@<hostname>:<port>
NO_PROXY=localhost,<pattern>,...
```

Note: Username and password may be omitted if not required.

3. Press *Ctrl + O*, then Enter to save and *Ctrl + X* to exit the text editor.
4. Source the environment file to apply modifications immediately.

```
$ source /etc/environment
```

5. Open *Settings* for Firefox or any other browser, and then click *Network*.

6. Under *Network Proxy*, set it to use system proxy or set it to manual and add the proxy with the same values as entered in the environment file.

2.3 Other Pre-requisites

The following prerequisites will need to be installed:

- Python* 3.10+
- Node.js 22.14.0+
- Graphics driver (https://github.com/intel/compute-runtime)
- NPU driver (https://github.com/intel/linux-npu-driver)

2.4 Application Ports

The following service ports will be required to be available before running the application:

Table 3. Application Port Numbers

Frontend	8080
Worker Services	5000-6000

3.0 Installation

This section provides streamlined, step-by-step instructions for installing and running the Edge AI Sizing Tool on a local or isolated system.

3.1 Step-by-Step Installation

Follow these steps to configure the environment and launch the application, allowing you to explore its features and capabilities with minimal setup:

1) Configure the system based on the hardware configuration & install the necessary drivers (Refer to Edge Developer Kit Reference Scripts).

2) Git clone the *Edge AI Sizing Tool* from the GitHub Repository to the system.

3) Navigate to the main directory of the folder.

4) Install all the dependencies, including required packages, for setting up *Python* and *NodeJS* environments.

```
$ sudo ./install.sh
```

5) Run the application.

```
$ ./start.sh
```

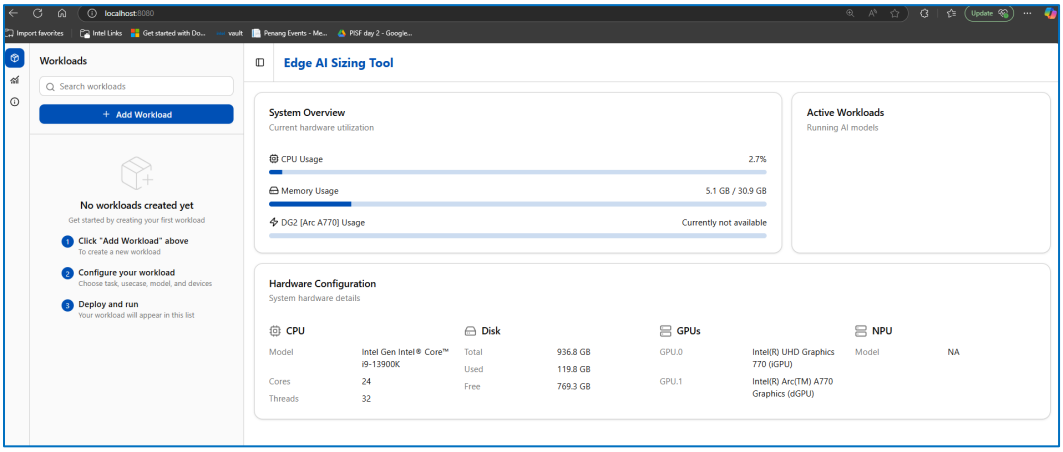
6) Open the http://localhost:8080 in your browser once the application has started.

7) Stop the application and all services, including the background workers.

```
$ ./stop.sh
```

The figure below shows the appearance of the initial UI when the application is launched and accessed via a web browser:

Figure 2. Edge AI Sizing Tool Main Dashboard



Disclaimer

1. Model Licensing

Before starting to run workloads in the application, please review the licensing agreements associated with each model. By proceeding with tasks and use cases, you acknowledge and agree to comply with the terms and conditions of the respective model licenses. Each model has its own licensing requirements, and it is your responsibility to ensure compliance. In the 5.0 Walkthrough for the supported models table section under each task and use cases, the official links will be provided where the models have been downloaded. You can review the licensing agreements through these links.

2. GStreamer* Licensing

Before starting to run Object Detection workload in the application, please review the licensing agreements associated with Gstreamer. By using GStreamer within the application, you acknowledge and agree to determine if your use requires any additional licenses. You are solely responsible for ensuring compliance with any licensing requirements associated with GStreamer*. In the 5.0 Walkthrough under the Object Detection workload, links to GStreamer* licensing information will be available for review.*

By running the application, you agree to adhere to the licensing terms of the models utilized within the tool and any requirements associated with GStreamer. Intel is not responsible for obtaining any such licenses or liable for any licensing fees due in connection with your use. In the 5.0 Walkthrough under the Object Detection workload, links to GStreamer licensing information will be available for review.*

4.0 Release Notes

Updated on: May 2025
Version: 2025.1 (Open-Source Release)

4.1 User Story, New Features and Enhancements

Table 4. User Story, New Features and Enhancements

Features	Summary
UI Revamp	Migrated the entire user interface to a whole new design.
Text to Image Use Case	Implemented the text-to-image generation function.
Automatic Speech Recognition Use Case	Implemented the speech transcription and translation function.
Object Detection Use Case	Implemented the real-time object detection (DLStreamer) function.
Text Generation Use Case	Implemented the chat generation function
System Information	Implemented displaying detailed system information.
Delete Workload	Implemented existing workload deletion.
Edit Workload	Modify the details of existing workloads.
OpenVINO™ Toolkit Auto Plugin	Implemented OpenVINO™ toolkit auto mode for multi-device selection. When more than one device is selected, the auto mode will utilize a priority list to manage device usage.
Backend Worker	Implemented status updates for workers that indicate online, offline, and model preparation states.
Automatic Speech Recognition Use Case	Added language selection as a configuration.
Automatic Speech Recognition Use Case	Added sample audio file.
Hardware Accelerator Model Name	Updated the GPU device identifiers (for example, GPU.0, GPU.1) to display descriptive names (for example, Intel® Arc™ Graphics, Intel® Arc™ A770 Graphics), based on device properties in Add, Edit, and Display Workloads.
Object Detection (DLStreamer) Use Case	Changed DLStreamer to utilize RTSP streaming that enables the stream to loop continuously.
Support OpenVINO Optimization	LLMs can now be exported using the optimum-cli which enables the download of OpenVINO-optimized LLMs from Hugging Face.

Simplify Quick Start	Simplified the quick start steps by creating shell and batch scripts.
Object Detection (DLStreamer) Use Case	Removed the segmentation models that were not functioning correctly for the object detection use case.
License Header	Added Apache 2.0 license.

4.2 Bug Fixes

Table 5. Bug Fixes

Features	Summary
System Overview	Fixed the issue preventing the display of system overview details.
System Monitoring	Fixed the issue where GPU utilization was returned as zero despite the GPU being running.
System Information	Fixed the issue where information is not displayed for dGPU and iGPU device when both are present.
System Monitoring	Fixed the issue where network requests continued running after being stopped, which caused memory leakage.
Text to Image Use Case	Fixed the issue of prompt button enabled while model is downloading.
Add & Edit Workload	Fixed the issue of adding or updating a workload fails if no device was selected.
Backend Worker	Fixed the issue where the worker failed to start if the user edited a workload to change the existing use case.
System Monitoring	Fixed the issue of an invalid device ID occurring when retrieving GPU utilization data.
Edit Workload	Fixed the issue of undefined devices error when editing a workload for the first time.
Automatic Speech Recognition use case	Fixed the issue preventing the Whisper model from running on the NPU.
Memory Usage	Fixed the issue with inaccurate data for memory usage.
Database Creation	Fixed the issue with failing to create database running “npm run demo”.
Edit Workload	Fixed the issue with latest information changes not being saved when the button is pressed.
PM2 process	Fixed the issue with PM2 process cannot restart after a workload was edited.
ESLint Configuration	Fixed the configuration error that previously allowed ESLint to pass during the build process without proper validation.

System Information	Fixed the “TypeError: speeds[i].toFixed” is not a function error.
FPS Chart	Fixed the issue where the chart was not updated.
Object Detection (DLStreamer) Use Case	Fixed the issue on UI not being able to respond when there is more than one use case running.
GPU Utilization Chart	Fixed the issue on utilization chart not being displayed for iGPU.
Object Detection (DLStreamer) Use Case	Fixed the issue with the use case being stuck at preparing workload dashboard.
Object Detection (DLStreamer) Use Case	Fixed the issue of clicking around the fields in Add Workload prevented the creation of a new workload.
System Monitoring	Fixed the issue where only one GPU utilization data was displayed in the monitor when multiple GPUs of the same model were present in the system.
Add & Edit Workload	Fixed the issue with camera input field not refreshing with list of available devices.
System Information	Fixed the issue on not displaying GPU device information.
Npm Run Demo Error	Fixed the issue where “run demo” failed to run.
ESlint Error	Fixed the ESlint errors that prevented “npm run build” from running successfully.
PM2 Process	Fixed the issue of PM2 process that did not restart after editing a workload.

4.3 Documentation

Table 6. Documentation

Features	Summary
Restructure	Cleaned and updated the content to reflect the new revamp changes.
Setup Platform	Updated the installation driver and configured the system link to use Edge Developer Kit Reference Scripts (DevKit).
Validated Hardware & Software	Added details of the validated hardware and software configuration details.
Application Ports	Included the application ports that should be available before running the application.
Run the Application	Added the quick-start application steps.
Deployment	Added steps on preparing and deploying the application.
Development	Added steps on setting up a development environment.

Troubleshooting	Updated guidance on how to debug and resolve certain known issues.
Limitations	Added current application limitations and known issues.
Disclaimer	Added Gstreamer licensing disclaimer and a hardware compatibility disclaimer.
Security Policy	Added security documentation.

§

5.0 Walkthrough

This section provides a description of Edge AI Sizing Tool, an explanation on how to use it and its features provided by this release. Below are the list of tasks and use cases provided in this release:

Table 7. Tasks and Use Cases

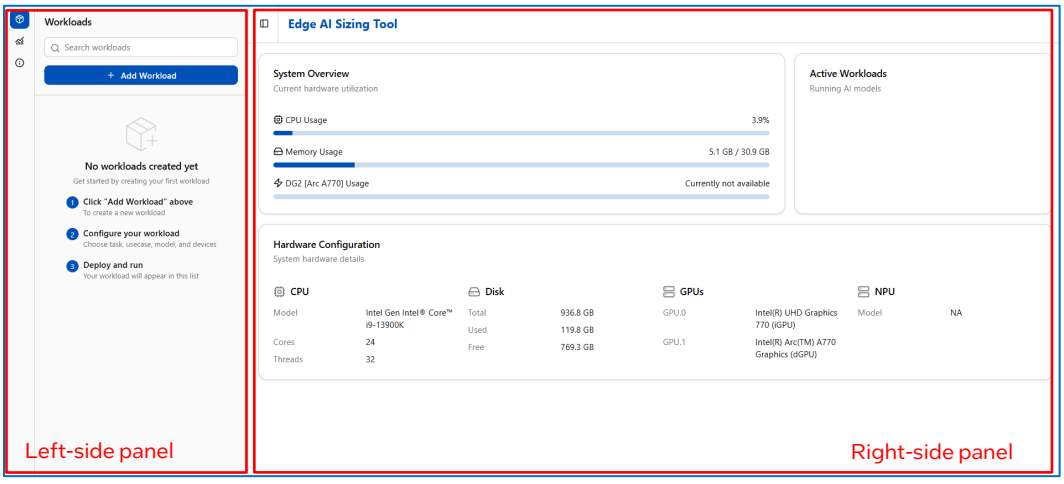
Computer Vision	• Object Detection • Text-to-Image
NLP	• Text Generation
Audio	• Automatic Speech Recognition

5.1 Main Dashboard

The new Edge AI Sizing Tool is divided into two sections:

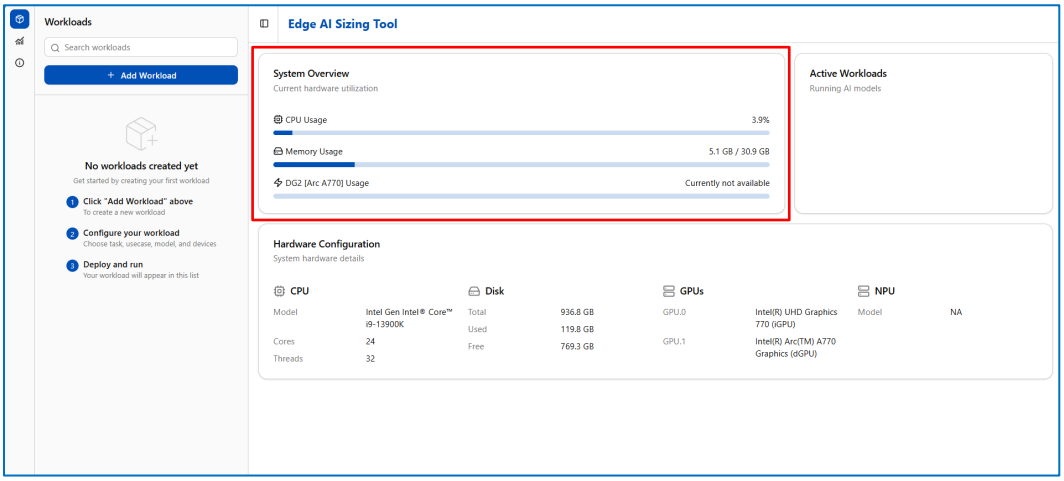
- **Left panel**, which includes the sidebar
- **Right panel**, which displays the main content

Figure 3. Edge AI Sizing Tool Main Dashboard



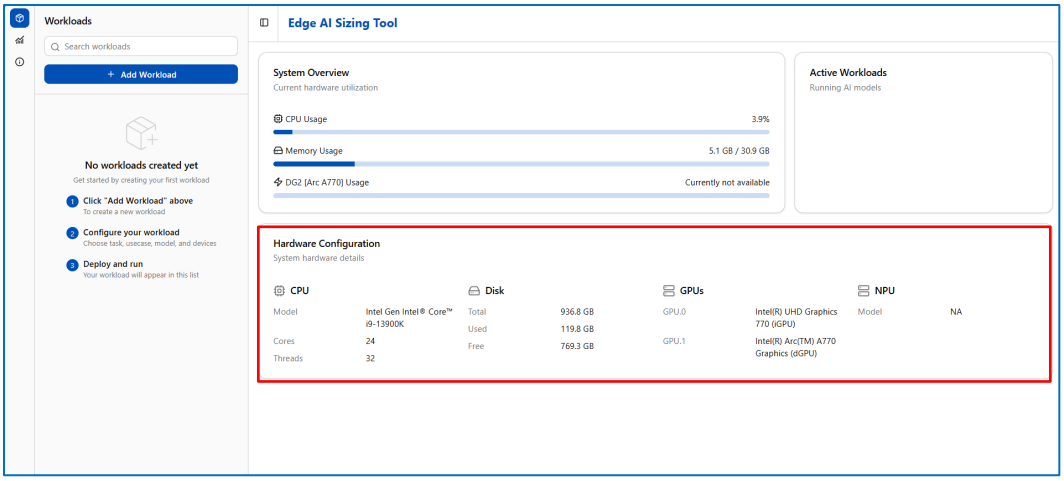
The main dashboard serves as the entry point and comprises a collection of integrated features that provide a comprehensive overview of the tool's offerings. The right panel includes a system overview section that displays hardware usage metrics, such as the current utilization of CPU, memory, iGPU, dGPU, and NPU.

Figure 4. System Overview



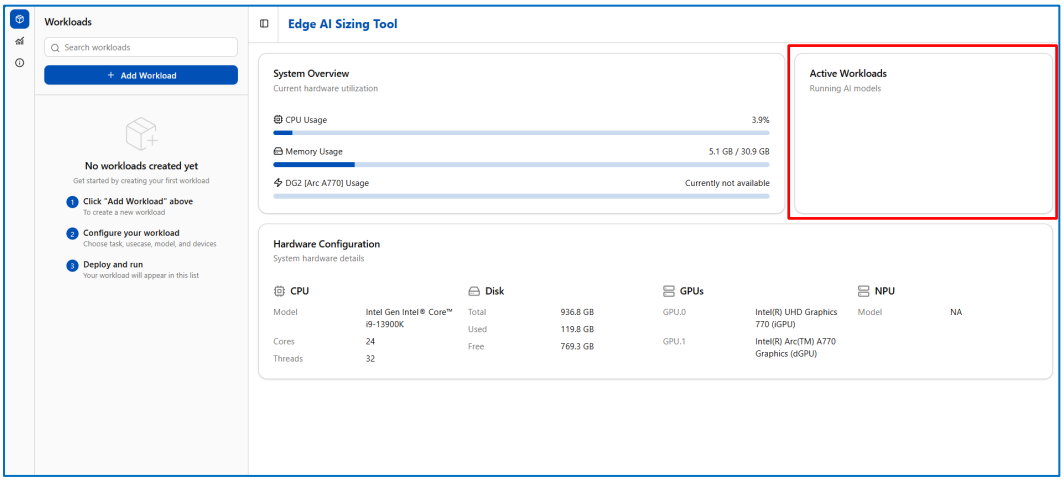
Below the system overview component, there is a hardware configuration section that provides detailed information about the system's hardware, including the CPU, disk space, GPUs, and NPU. For the CPU, it displays the name, number of cores, and threads. Regarding disk space, it shows the total capacity, the amount utilized, and the available space. For GPUs and the NPU, it presents the model names.

Figure 5. Hardware Configuration



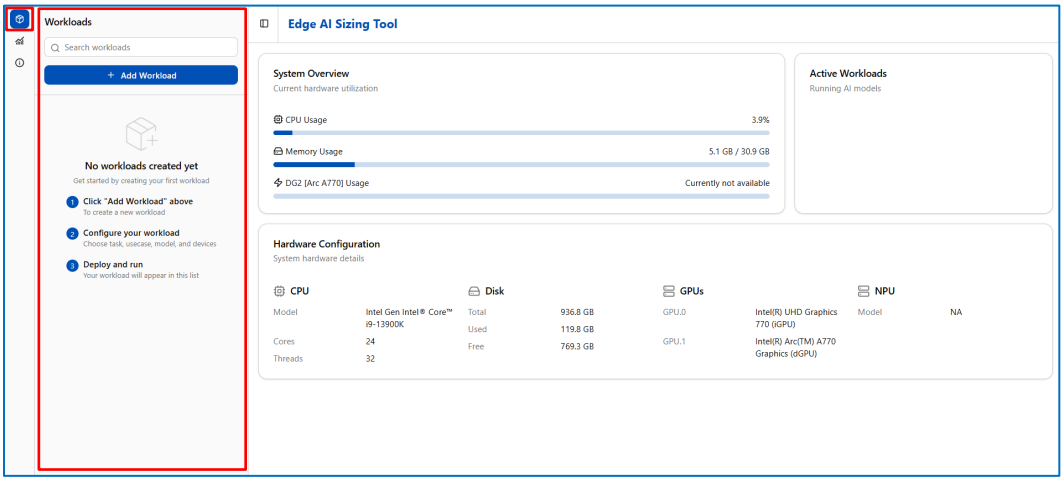
On the right side of the system overview component, there is an active workload section that provides a concise summary of the currently active and running AI workloads, including details such as the model, use case, and accelerator.

Figure 6. Active Workloads under Main Dashboard



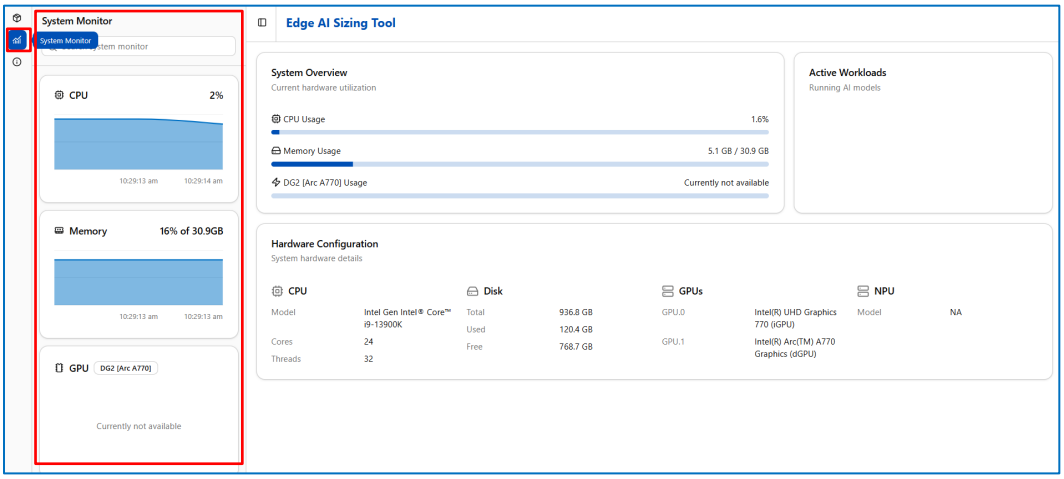
At the top left-most of the left panel, there are three icon buttons. Clicking the first button opens the workload management feature in the sidebar, located in the left panel, which allows users to view, create, edit, and delete AI workloads. All the previously saved workloads will be listed in this sidebar, and selecting a specific workload will display its content in the right panel.

Figure 7. Workloads



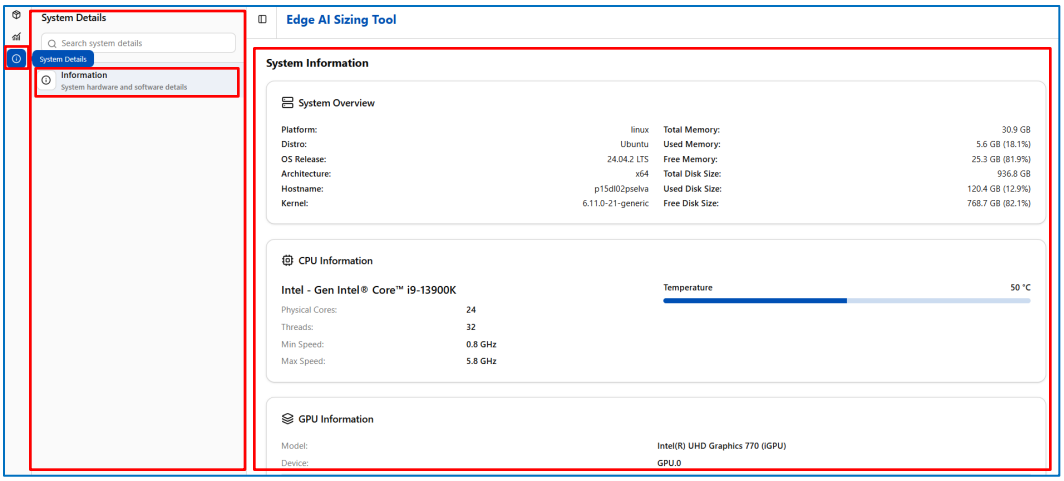
Clicking the second button opens the system monitor feature, which provides real-time visualization of CPU, iGPU, dGPU, NPU, and memory usage in the sidebar of the left panel. This feature is particularly useful for monitoring system performance while running workloads.

Figure 8. System Monitor



Clicking the third button opens the Information section in the sidebar of the left panel. Selecting Information from the sidebar will display detailed system information in the right panel, including comprehensive hardware and software details such as platform, distribution, OS release, architecture, host name, kernel, total memory, used memory, free memory, total disk size, used disk size, free disk size, CPU model, physical cores, threads, minimum and maximum speeds, temperature, GPU model and device, and NPU model.

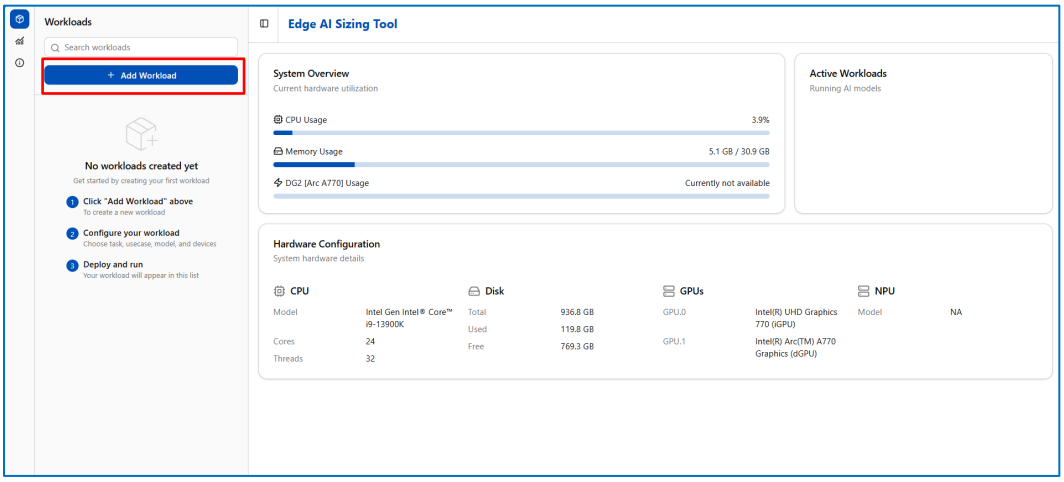
Figure 9. System Information Dashboard



5.2 Add a Workload

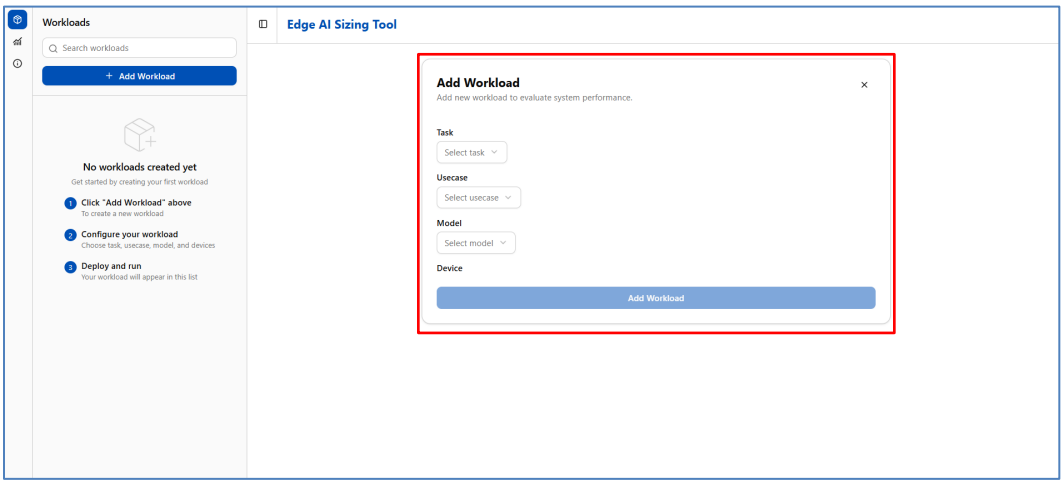
To create a new AI workload for evaluating the system performance, click on the first button icon on the left-most under the left-side panel.

Figure 10. Add Workload Button



Next, click the Add Workload button to open a form in the right panel for adding a new workload. All fields in the form must be filled out. Once completed, click the Add Workload button to create the new AI workload.

Figure 11. Add Workload Form



Under the workload form, there will be three tasks available:

- Computer Vision
- Natural Language Processing
- Audio

For **Computer Vision**, there will be two use cases available:

- Object Detection
- Text-to-Image

For **NLP**, there will be a single use case:

- Text Generation

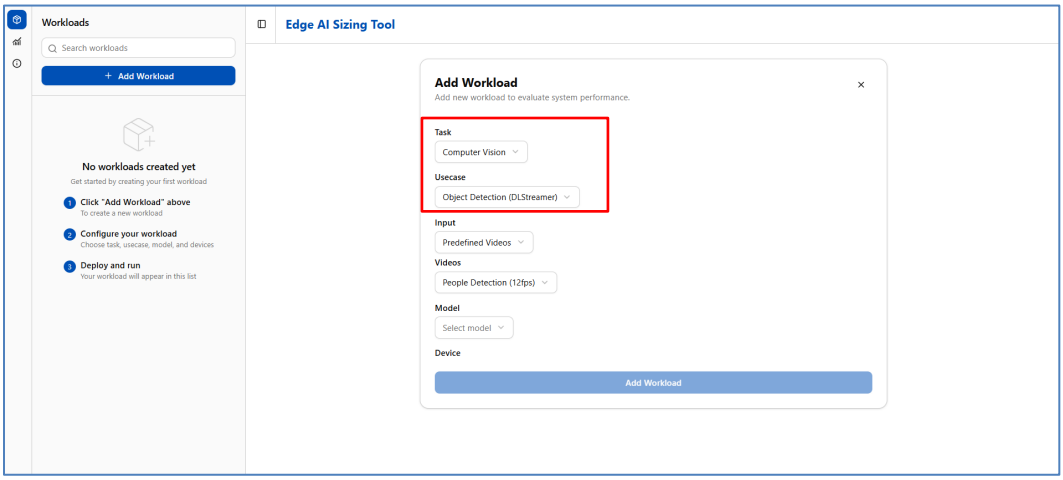
For **Audio**, there will be a single use case:

- NLP

5.3 Computer Vision (Object Detection)

To create a workload using the Object Detection use case, select Computer Vision in the Task field and Object Detection (DLStreamer) in the Use Case field. Please remember to review the licensing agreements as outlined in the Disclaimer under 3.0 Installation.

Figure 12. Task & Use Case Fields for Computer Vision (Object Detection)

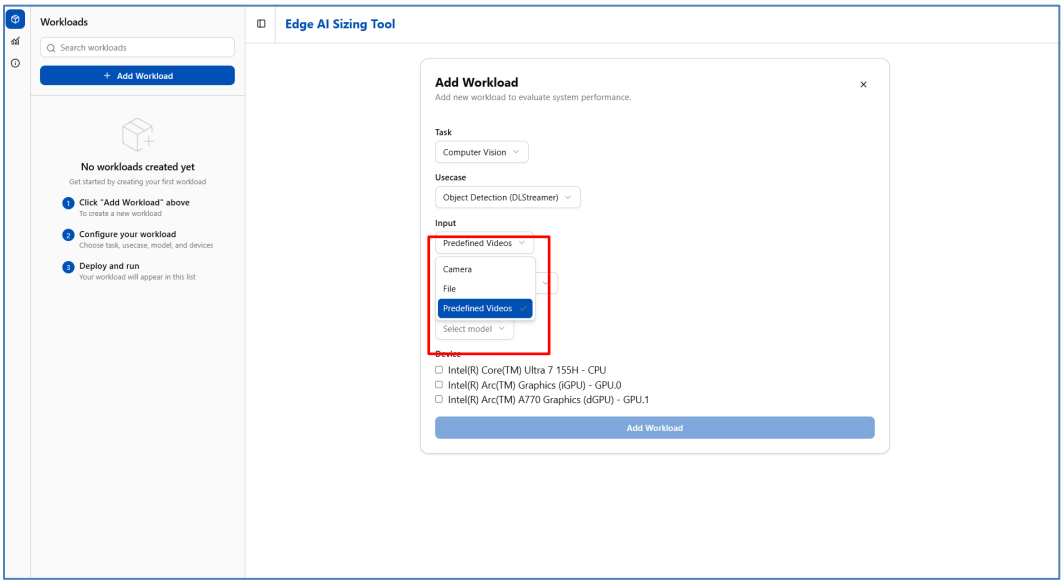


In the Input field, there will be three options:

- Camera
- File
- Predefined Videos

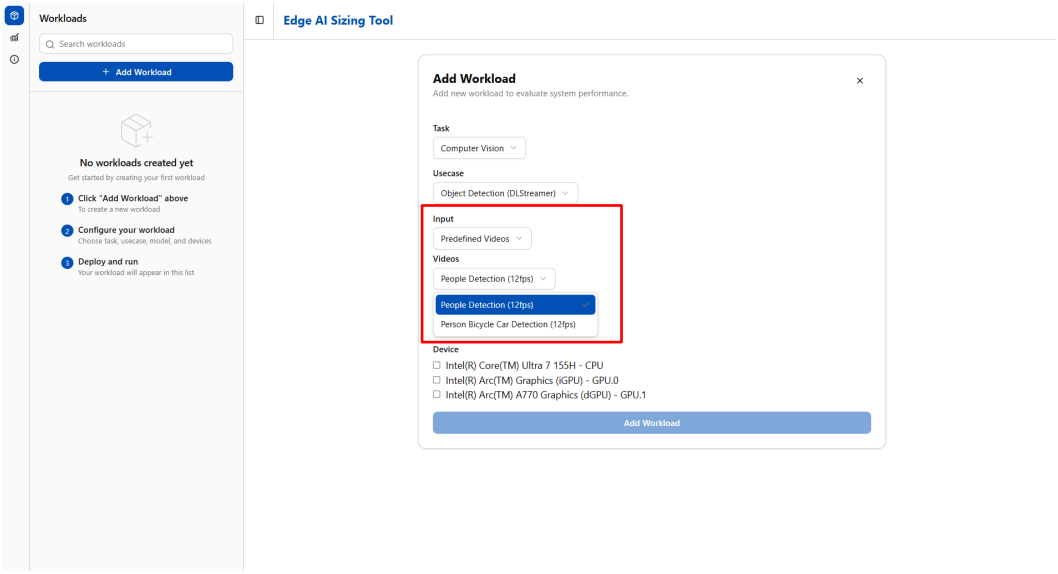
Depending on the selection, an additional specific field will be available to accommodate the chosen input type.

Figure 13. Input Field for Computer Vision (Object Detection)



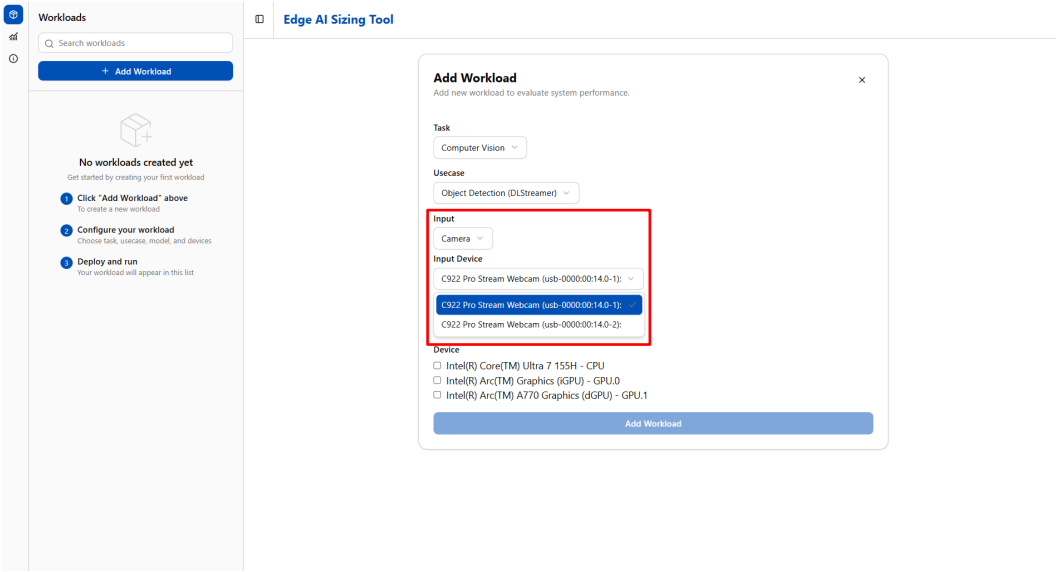
If **Predefined Videos** is selected as the input, a Videos field will appear, which enables to choose from a list of videos predefined by the tool for use in object detection.

Figure 14. Input (Predefined Videos) and Videos Fields for Computer Vision (Object Detection)



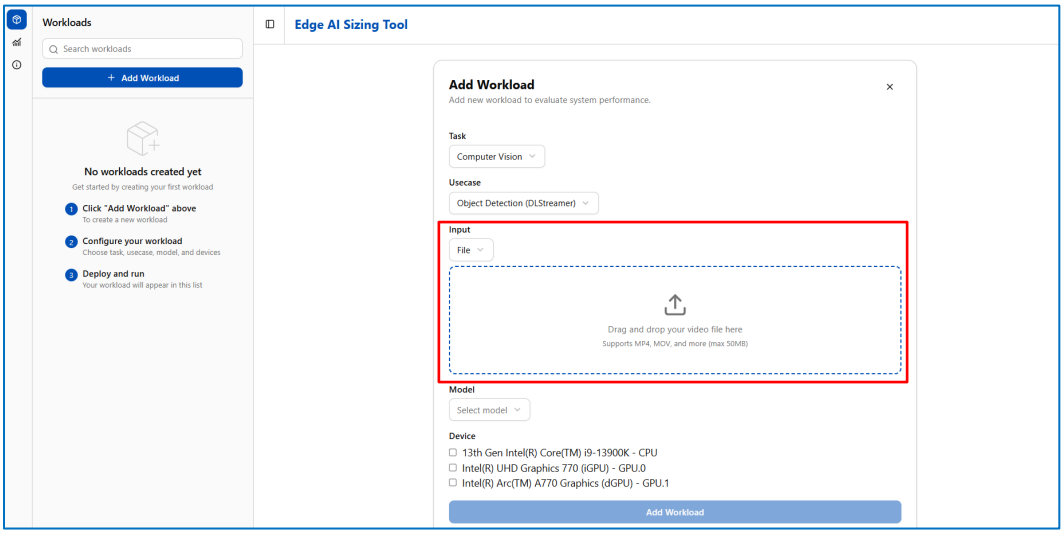
If **Camera** is selected as the input, an Input Device field will appear, where there need to select one of the camera devices connected to the system for use in object detection.

Figure 15. Input (Camera) and Input Device Fields for Computer Vision (Object Detection)



If **File** is selected as the input, a Video File Upload field will appear, allowing you to select a video from your system to be used for object detection.

Figure 16. Input (File) and Video File Drop Fields for Computer Vision (Object Detection)



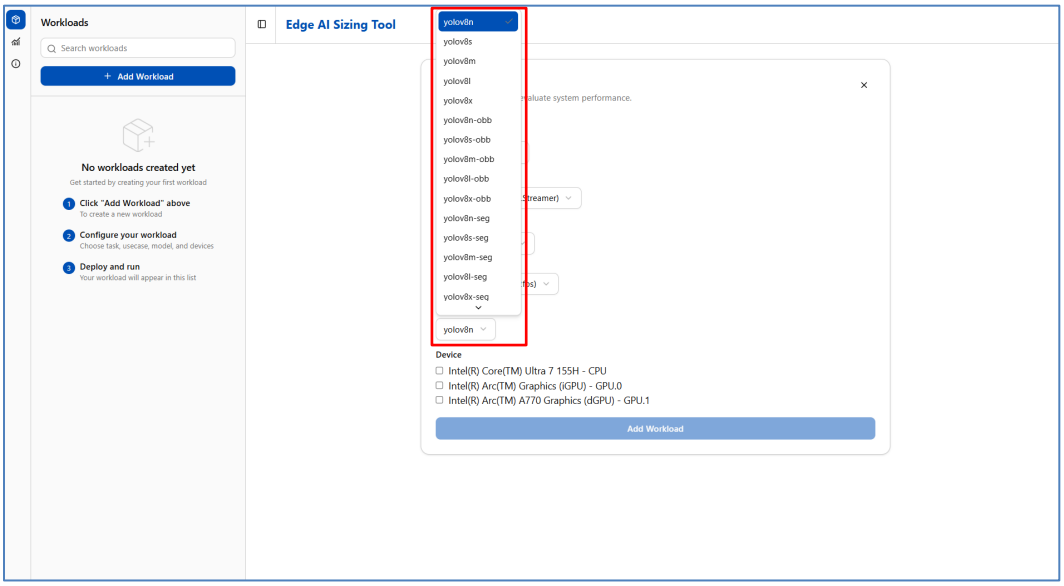
In the Model field, a list of supported and available AI models for object detection is provided. The following OpenVINO™ toolkit supported models are available for selection in the Object Detection (DL Streamer) use case, along with links to where they can be downloaded.

Table 8. Object Detection Supported Models

yolo8	yolo8n, yolo8s, yolov8m, yolov8l, yolov8x, yolov8n-obb, yolov8s-obb, yolov8m-obb, yolov8l-obb, yolov8x-obb	Ultralytics Docs
yolo9	yolov9t, yolov9s, yolov9m, yolov9c, yolov9e	
yolo10	yolov10n, yolov10s, yolov10m, yolov10b, yolov10l, yolov10x,	
yolo11	yolo11n, yolo11s, yolo11m, yolo11l, yolo11x, yolo11n-obb, yolo11s-obb, yolo11ms-obb, yolo11l-obb, yolo11x-obb, yolo11n-pose, yolo11s-pose, yolo11m-pos, yolo11l-pose, yolo11x-pose	

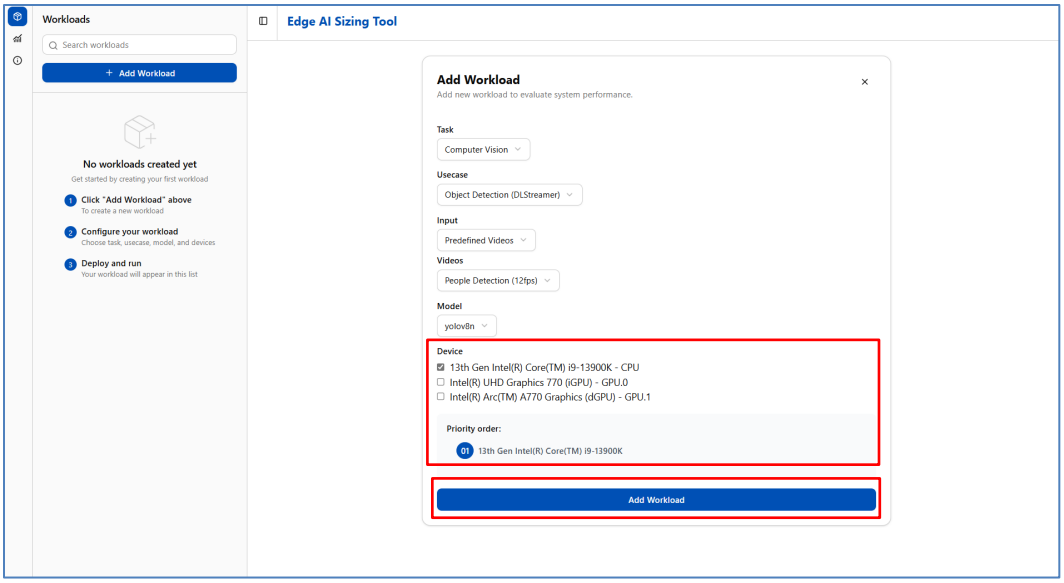
Select the model that best suits the object detection requirements, but remember to review the licensing agreements as outlined in the *Disclaimer* under 3.0 Installation.

Figure 17. Model Field for Computer Vision (Object Detection)



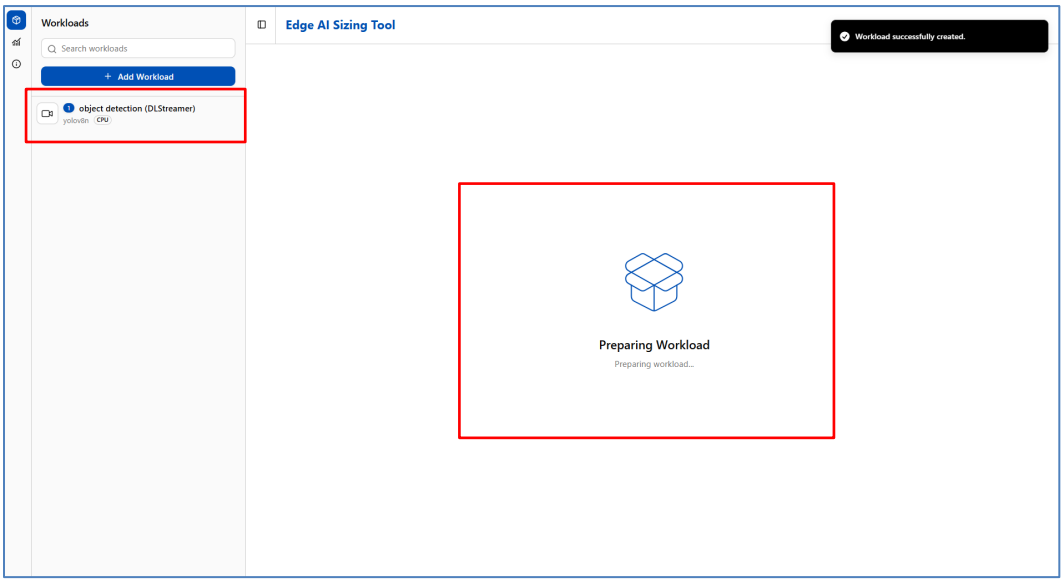
In the Device field, a list of available hardware accelerators is provided, from which a single accelerator must be selected. Once all fields are completed, the Add Workload button will be enabled, and clicking it will save the new workload.

Figure 18. Device Field & Add Workload Button for Computer Vision (Object Detection)



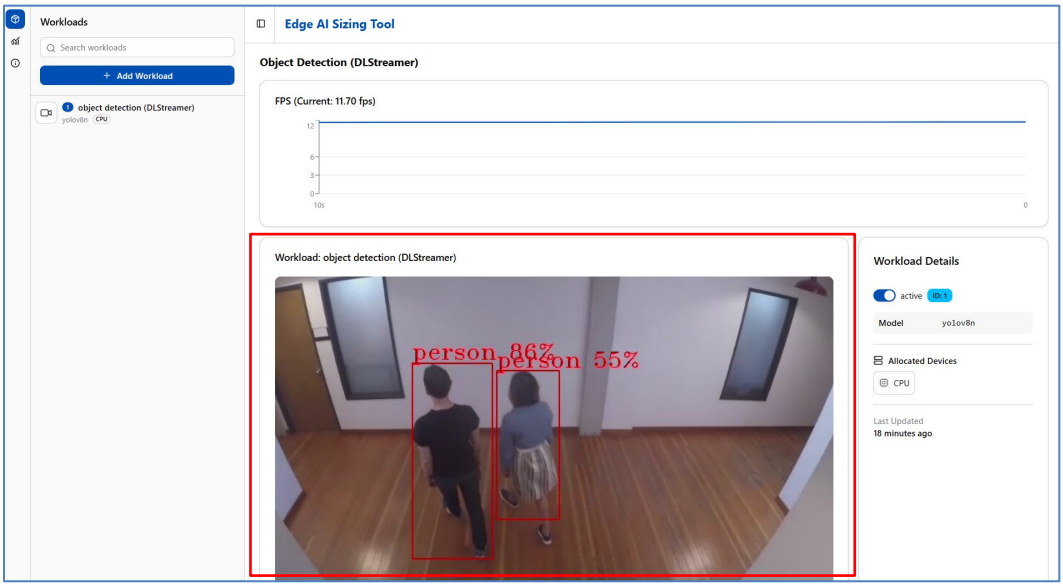
After clicking Add Workload, the right panel indicates that the tool is preparing the workload. This process may take some time, depending on the speed of the internet connection, as the model will be downloaded in the background if it hasn't been downloaded previously. In the sidebar of the left panel, a new Object Detection (DLStreamer) workload will be listed, which can be selected to view the workload content in the right panel.

Figure 19. Preparing Workload Dashboard for Computer Vision (Object Detection)



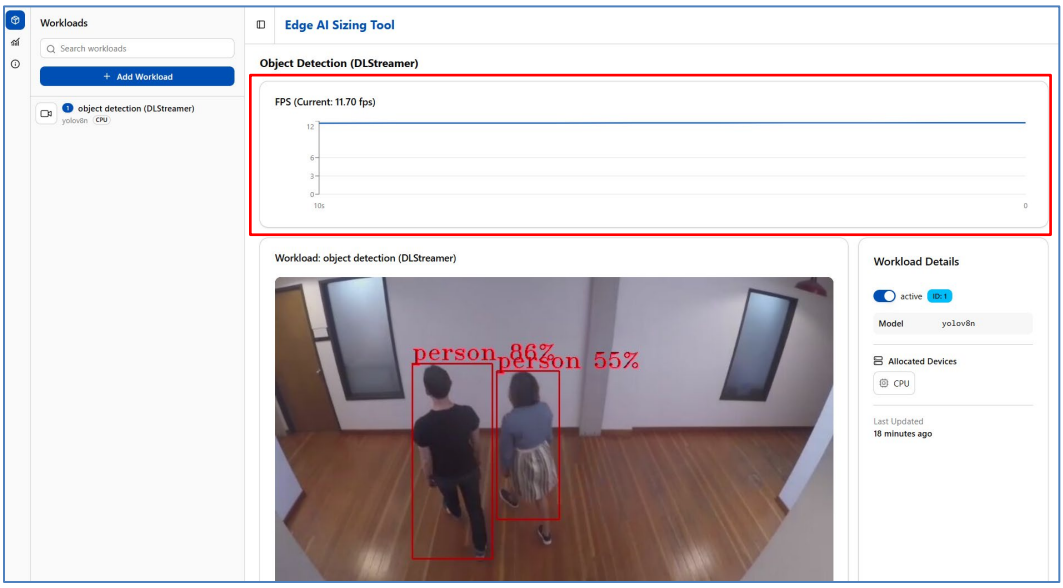
Upon successful completion of the Object Detection workload preparation process, the workload content will be displayed in the right panel. The object detection stream will automatically begin running based on the selected video input, in this case, the predefined video.

Figure 20. Computer Vision (Object Detection) Workload Dashboard



During the running object detection process, additional information, such as FPS, is provided, indicating the number of frames processed per second in the video. This information can be observed at the top of the right panel.

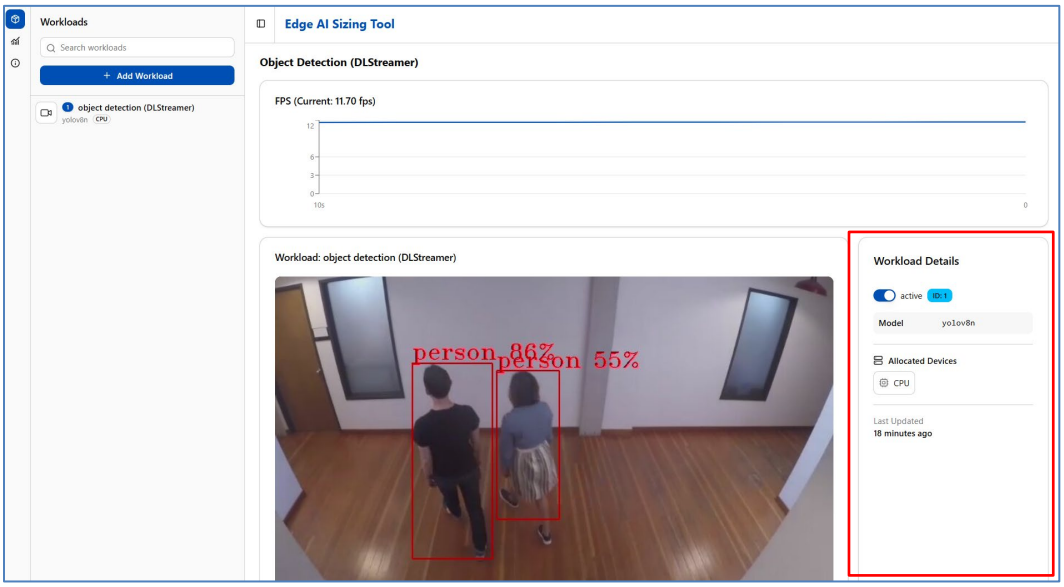
Figure 21. FPS Graph for Computer Vision (Object Detection)



In addition to the object detection stream, the right panel also displays details of the Object Detection workload, including the workload ID, status (active or inactive),

model name, allocated accelerator devices, and the last update timestamp. An active toggle button is available to disable the Text to Image workload, which prevents it from being active and running in the system.

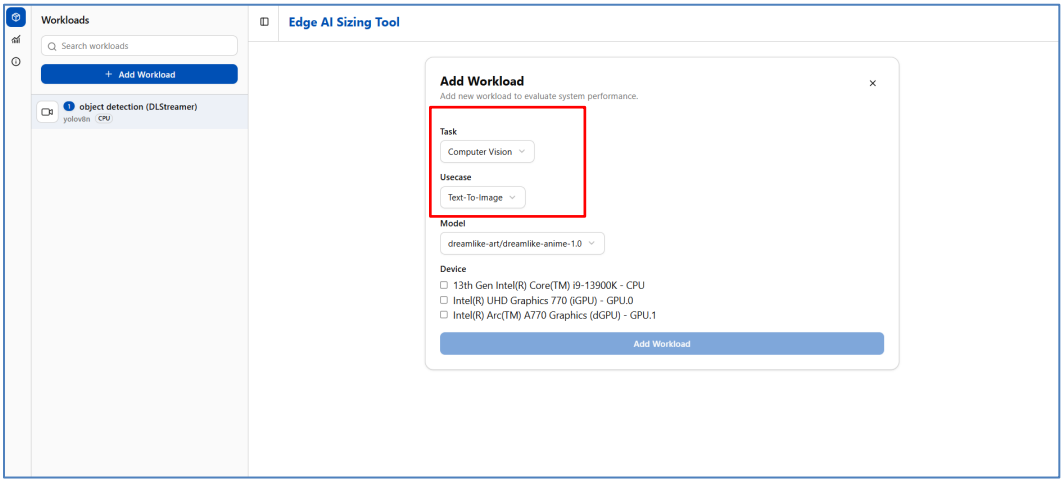
Figure 22. Workload Details for Computer Vision (Object Detection)



5.4 Computer Vision (Text to Image)

To create a workload using the Text-To-Image use case, Computer Vision needs to be selected under the Task field, and Text-To-Image needs to be selected under the Use Case field.

Figure 23. Task & Use Case Fields for Computer Vision (Text to Image)



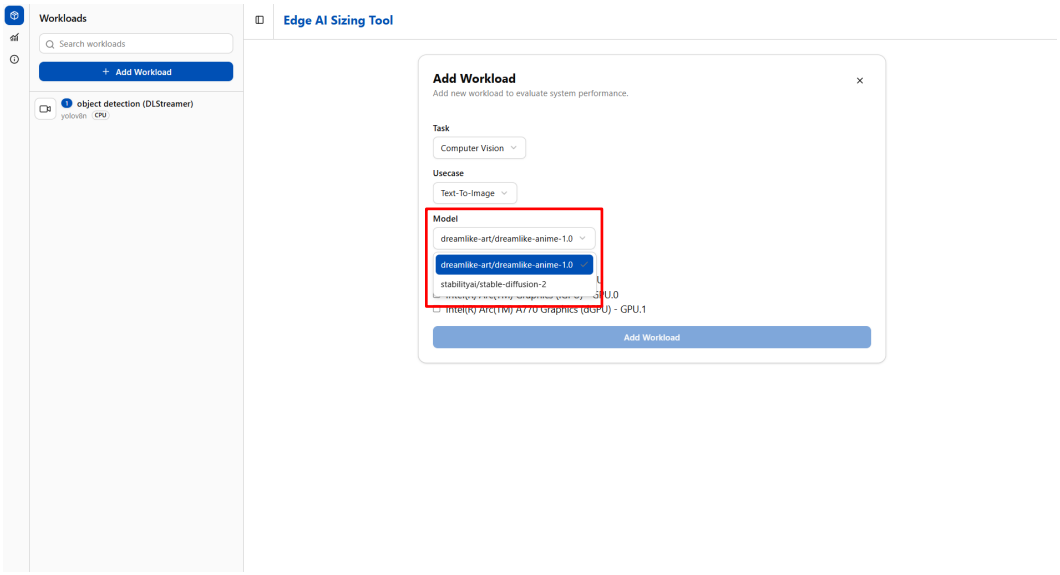
In the Model field, a list of supported and available AI models for Text-to-Image generation is provided. One of these models must be selected for image generation. The following OpenVINO™ toolkit supported models are available for selection in the Text-to-Image use case, along with links to where they can be downloaded:

Table 9. Text to Image Supported Models

dreamlike-art	dreamlike-anime-1.0	Hugging Face
stabilityai	stable-diffusion-2	Hugging Face

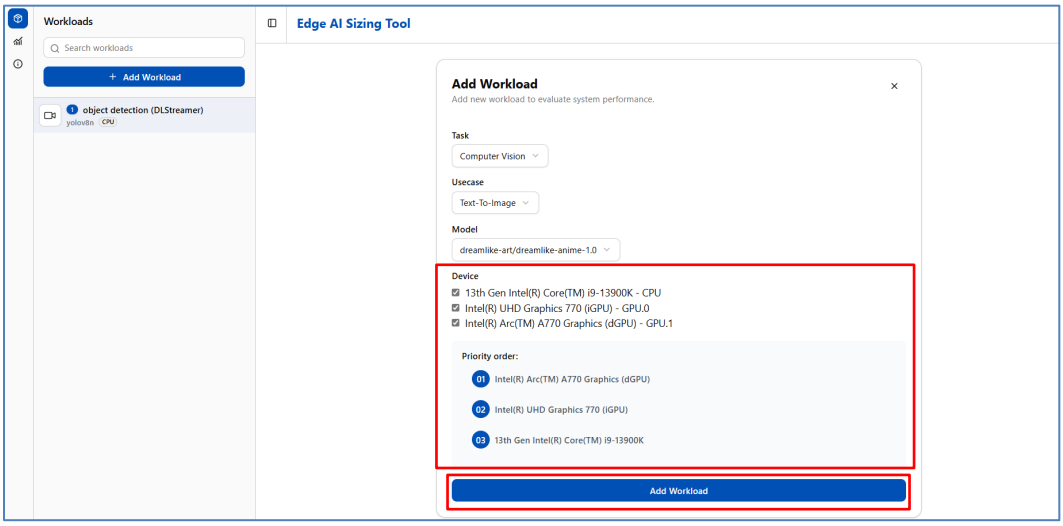
Select the model that best fits the image generation requirements, but remember to review the licensing agreements as outlined in the *Disclaimer* under 3.0 Installation.

Figure 24. Model Field for Computer Vision (Text to Image)



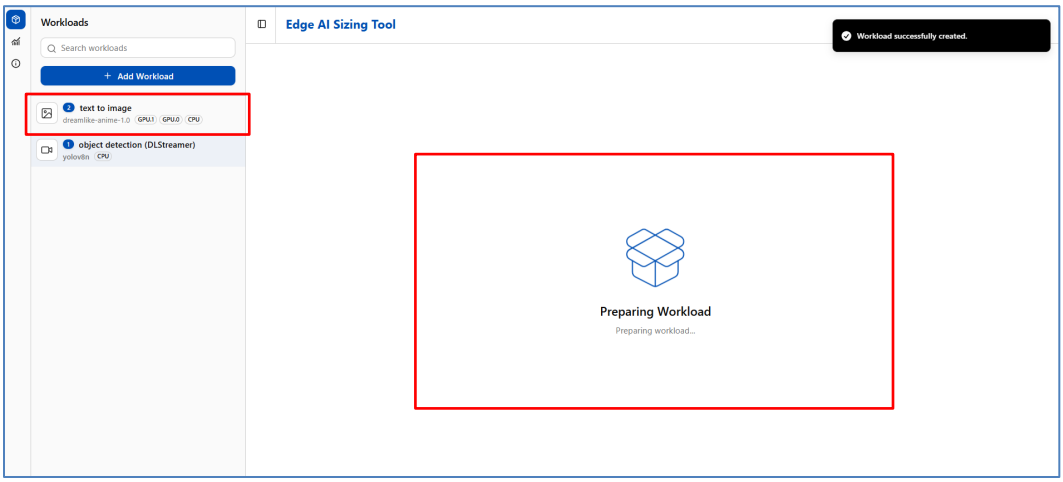
In the Device field, one or more hardware accelerators can be selected to execute the workload. If multiple accelerators are chosen, a priority order list will be displayed, indicating that the workload will be executed based on this order. If the primary accelerator device is busy, the next device in the list will be used. After completing all fields, click Add Workload to save the workload.

Figure 25. Device Field & Add Workload Button for Computer Vision (Text to Image)



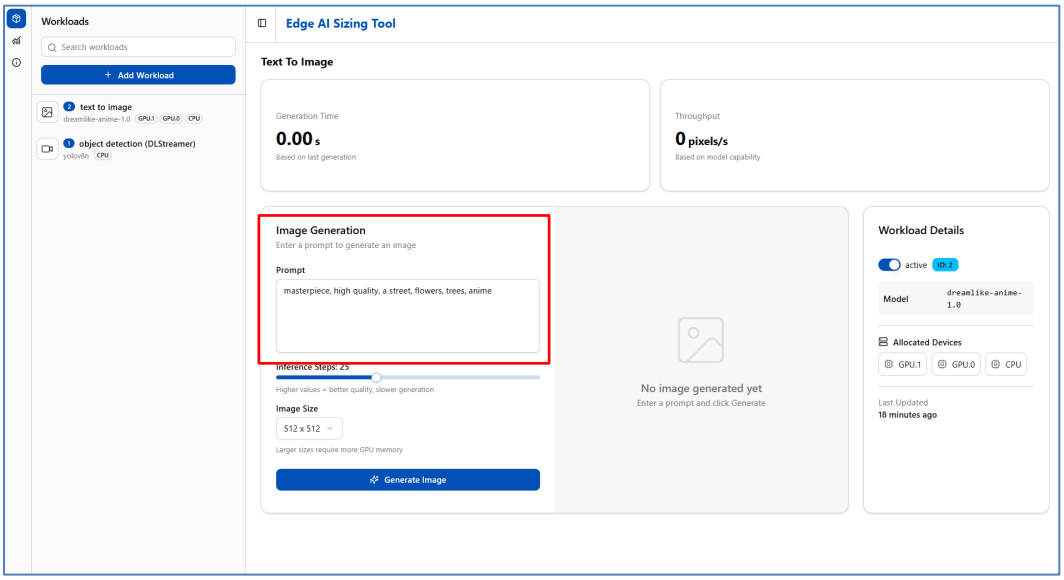
After clicking Add Workload, the right panel indicates that the tool is preparing the workload. This process may take some time, depending on the speed of the internet connection, as the model will be downloaded in the background if it hasn't been downloaded previously. In the sidebar of the left panel, a new Text-To-Image workload will be listed, which can be selected to view the workload content in the right panel.

Figure 26. Preparing Workload Dashboard for Computer Vision (Text to Image)



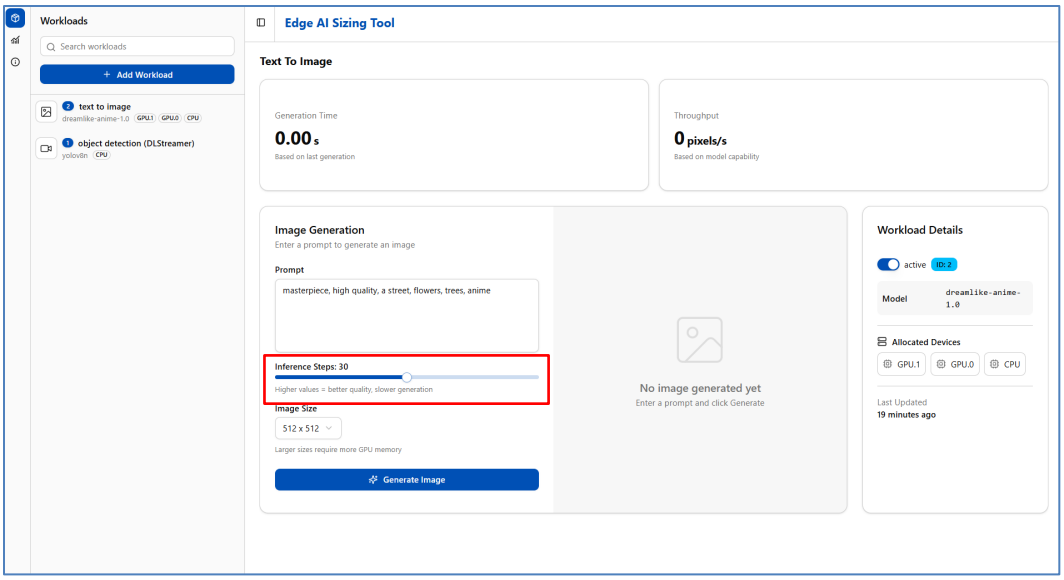
Once the Text-To-Image workload preparation process is successfully completed, the workload content will be displayed in the right panel. Before executing the Text-To-Image workload, several settings can be adjusted using the provided configuration form. In the Prompt field, the prompt text can be modified, or the default text can be used for testing purposes.

Figure 27. Prompt Field for Computer Vision (Text to Image)



The inference steps have a default value of 25, which can be adjusted to determine how many times the model will perform inferencing based on the provided prompt text. Increasing the number of inference steps generally results in higher image quality but also leads to slower generation times.

Figure 28. Inference Steps Field for Computer Vision (Text to Image)

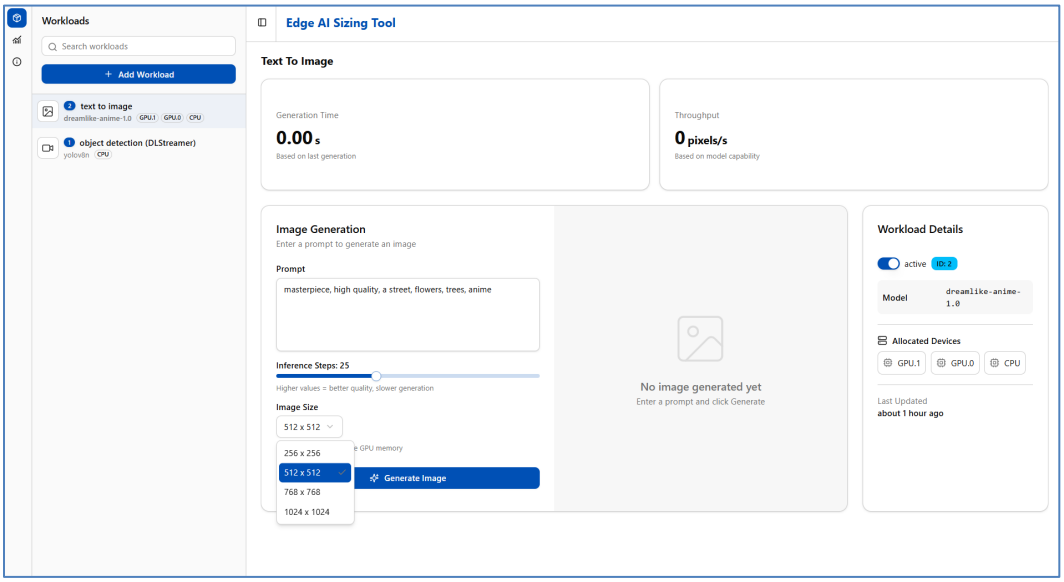


For the Image Size, several options are available to specify the dimensions of the generated image:

- 512x512 (default size)
- 256x256
- 768x768
- 1024x1024

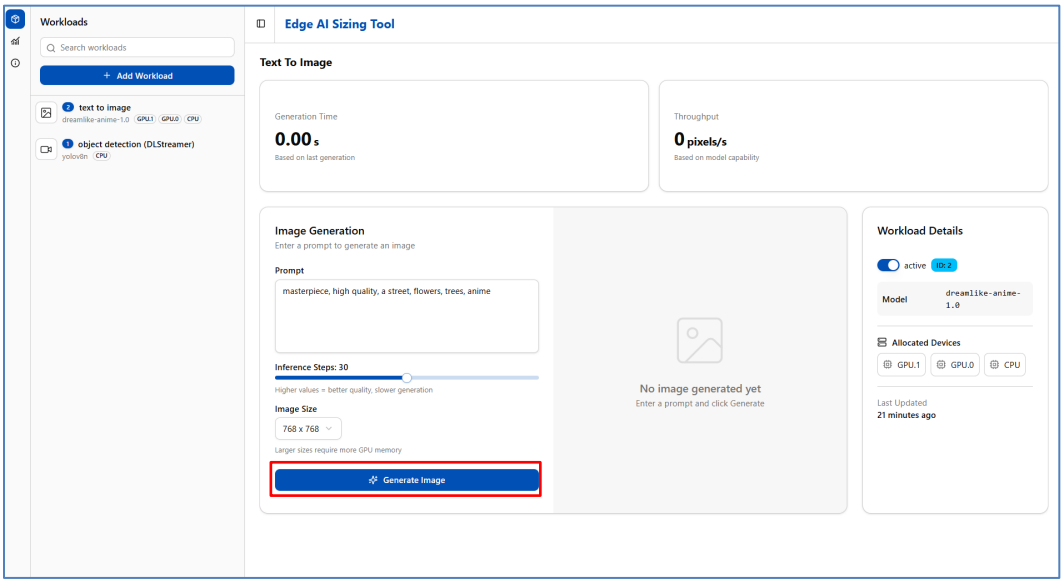
Larger image sizes require more memory for processing.

Figure 29. Image Size Field for Computer Vision (Text to Image)



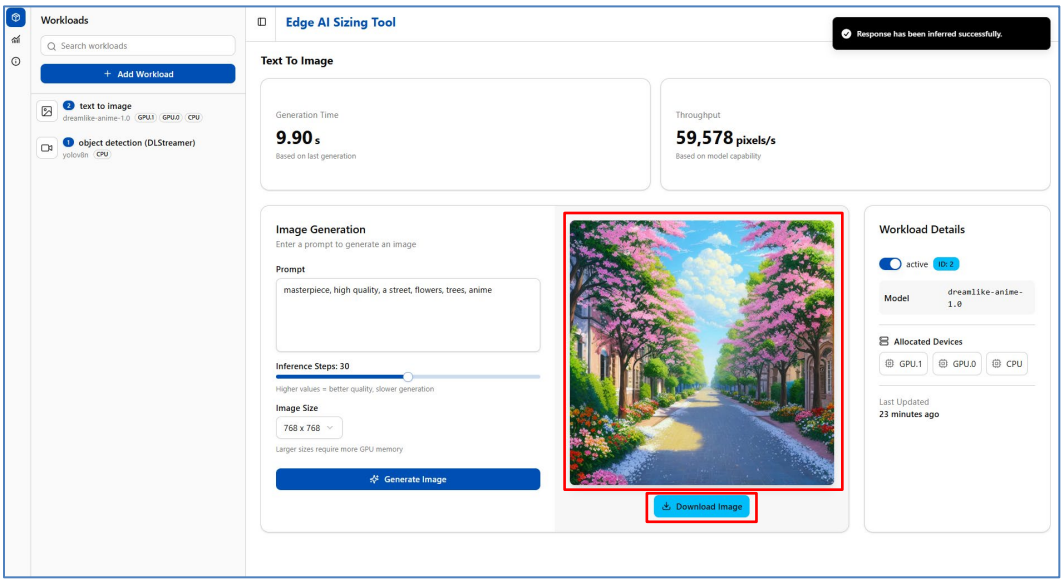
Once the configuration adjustments are complete according to testing requirements, click Generate Image to initiate the inferencing process for image generation.

Figure 30. Generate Image Button for Computer Vision (Text to Image)



Upon successful completion of the inference, the generated image will be displayed in the image generation placeholder next to the configuration form. Additionally, a Download Image button will be available that enables the generated image to be downloaded to the system when clicked.

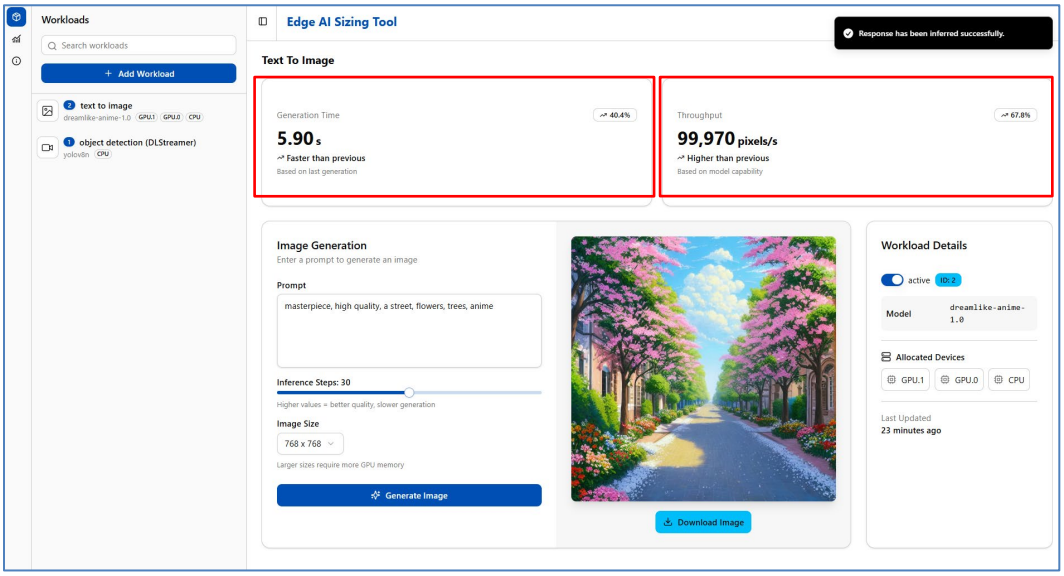
Figure 31. Generated Image Placeholder & Download Button for Computer Vision (Text to Image)



In addition to the generated image, additional information is provided, such as the generation time, which indicates the time taken in seconds to complete the

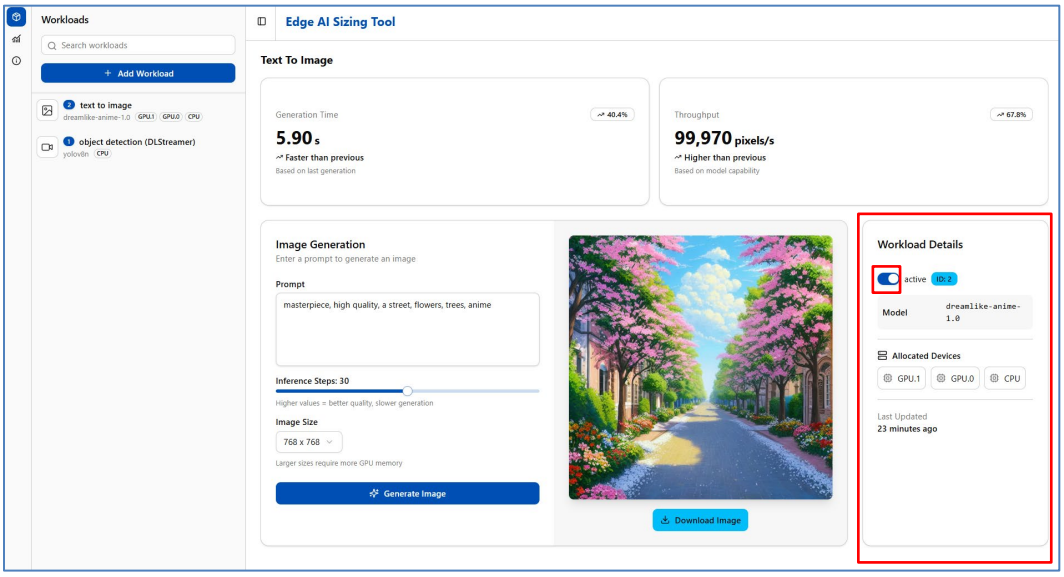
inferencing during the last generation. The throughput value, which reflects the number of pixels processed based on the model's capacity, will also be displayed. This information can be observed at the top of the right panel.

Figure 32. Generation Time & Throughput for Computer Vision (Text to Image)



In addition to the image generation placeholder, the right panel displays Text-to-Image workload details, including the workload ID, status (active or inactive), model name, allocated accelerator devices, and the last update timestamp. An active toggle button is available to disable the Text-to-Image workload, preventing it from running in the system.

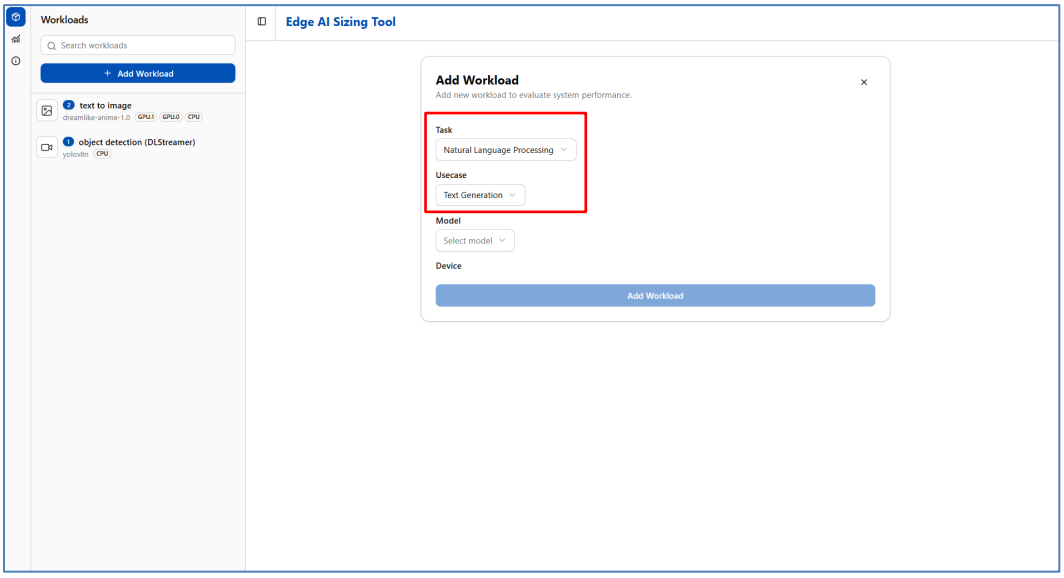
Figure 33. Workload Details for Computer Vision (Text to Image)



5.5 Text Generation (NLP)

To create a workload using the NLP use case, select "Text Generation" in the Task field and Natural Language Processing in the Use Case field.

Figure 34. Task & Use Case Fields for Text Generation (NLP)



In the Model field, a list of supported and available AI models for NLP is provided. One of these models must be selected for Text Generation. The following OpenVINO™

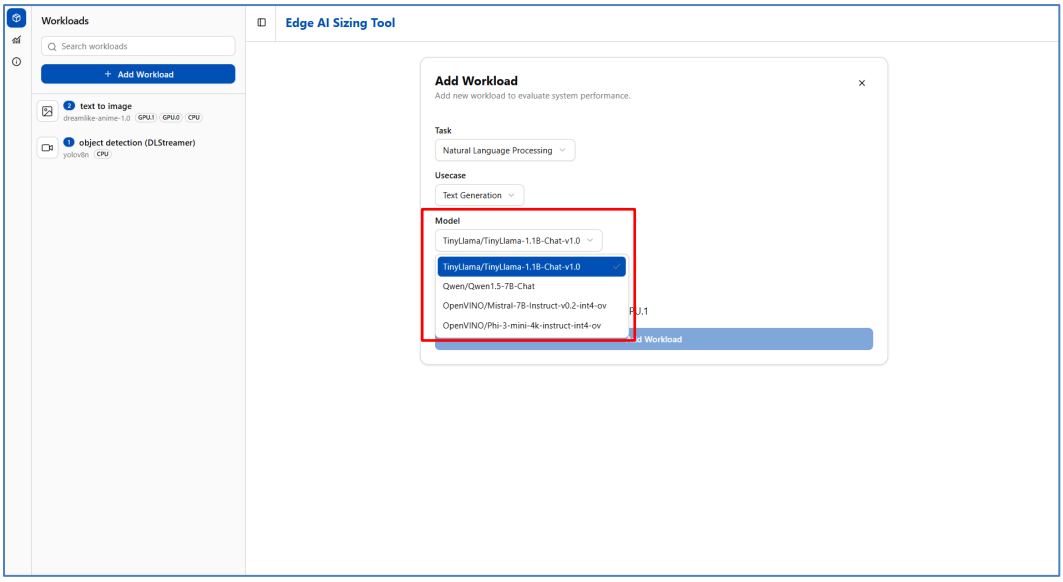
supported models are available for selection in the Text Generation use case, along with links to where they can be downloaded:

Table 10. Text Generation Supported Models

TinyLlama	TinyLlama-1.1B-Chat-v1.0	Hugging Face
Qwen	Qwen1.5-7b-Chat	Hugging Face
OpenVINO	Mistral-7B-Instruct-v0.2-int4-ov	Hugging Face
	Phi-3-mini-4k-instruct-int4-ov	Hugging Face

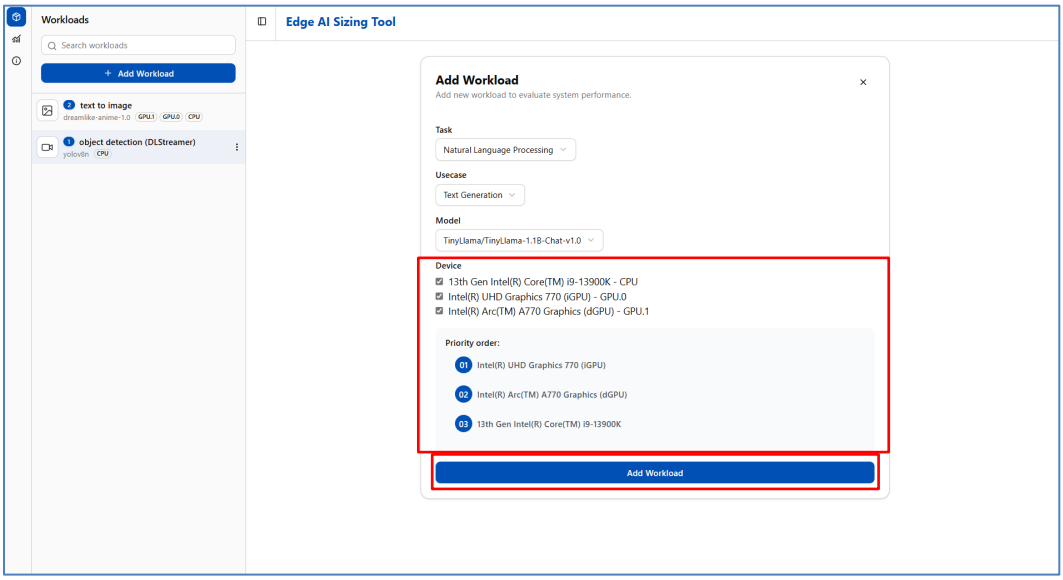
Select the model that best fits the text generation requirements, but please remember to review the licensing agreements as outlined in the *Disclaimer* under 3.0 Installation.

Figure 35. Model Field for Text Generation (NLP)



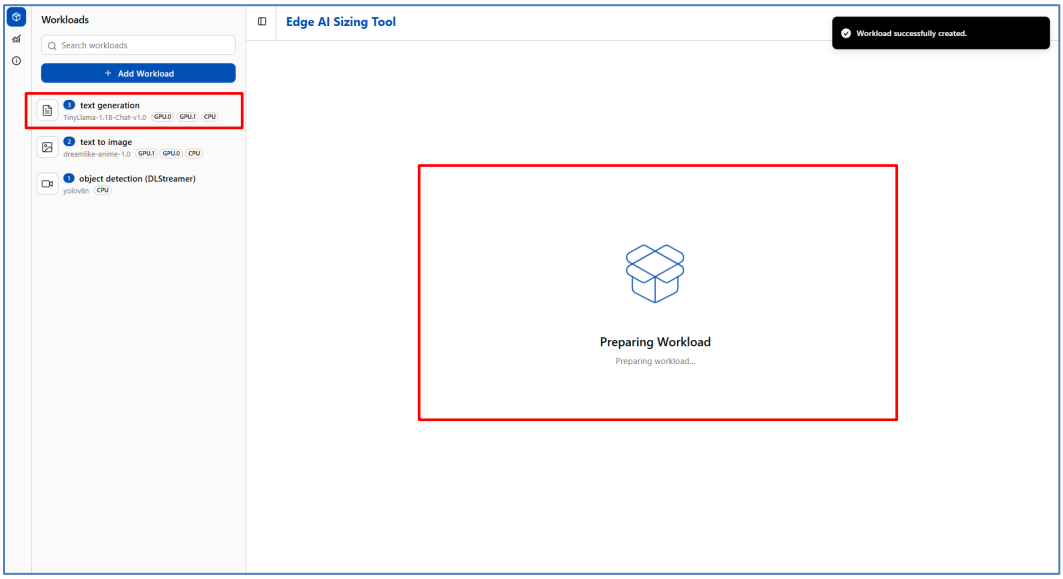
In the Device field, one or more hardware accelerators can be selected to execute the workload. If multiple accelerators are chosen, a priority order list will be displayed, indicating that the workload will be executed based on this order. If the primary accelerator device is busy, the next device in the list will be used. After completing all fields, click Add Workload to save the workload.

Figure 36. Device Field & Add Workload Button for Text Generation (NLP)



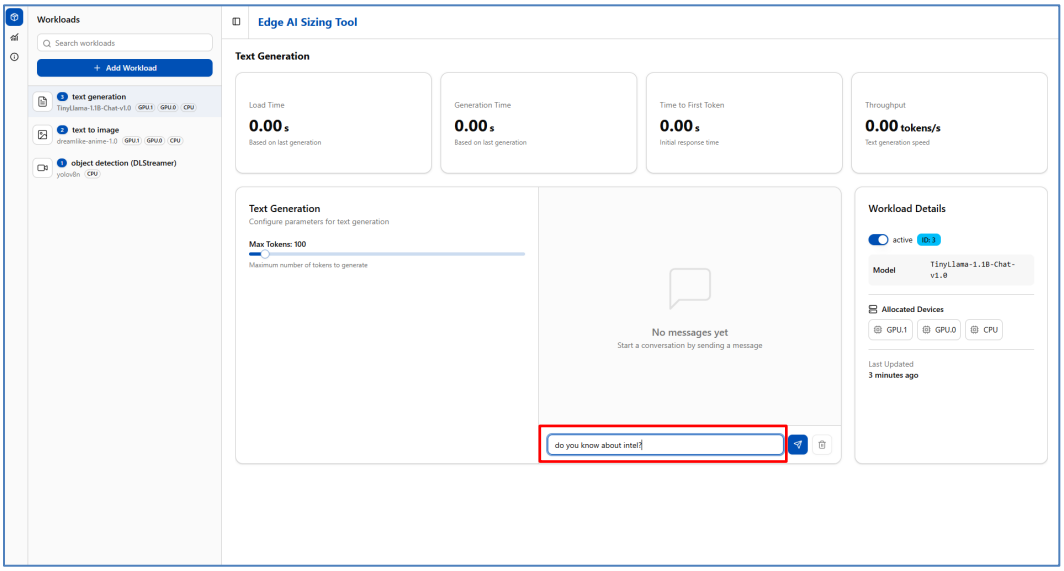
After clicking Add Workload, the right panel indicates that the tool is preparing the workload. This process may take some time, depending on the speed of the internet connection, as the model will be downloaded in the background if it hasn't been downloaded previously. In the sidebar of the left panel, a new Text Generation workload will be listed, which must be clicked to view the workload content in the right panel.

Figure 37. Preparing Workload Dashboard for Text Generation (NLP)



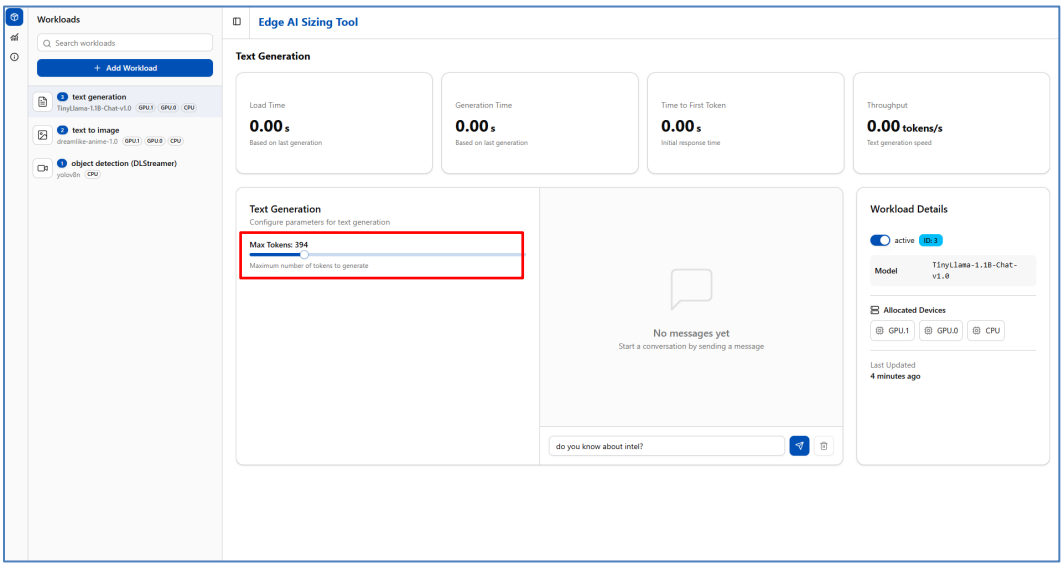
Once the Text Generation workload preparation process is successfully completed, the workload content will be displayed in the right panel. To execute inferencing using the Text Generation workload, the Type Your Message field must be filled with the desired questions or queries, which will be used by the selected model for text completion.

Figure 38. Message Field for Text Generation (NLP)



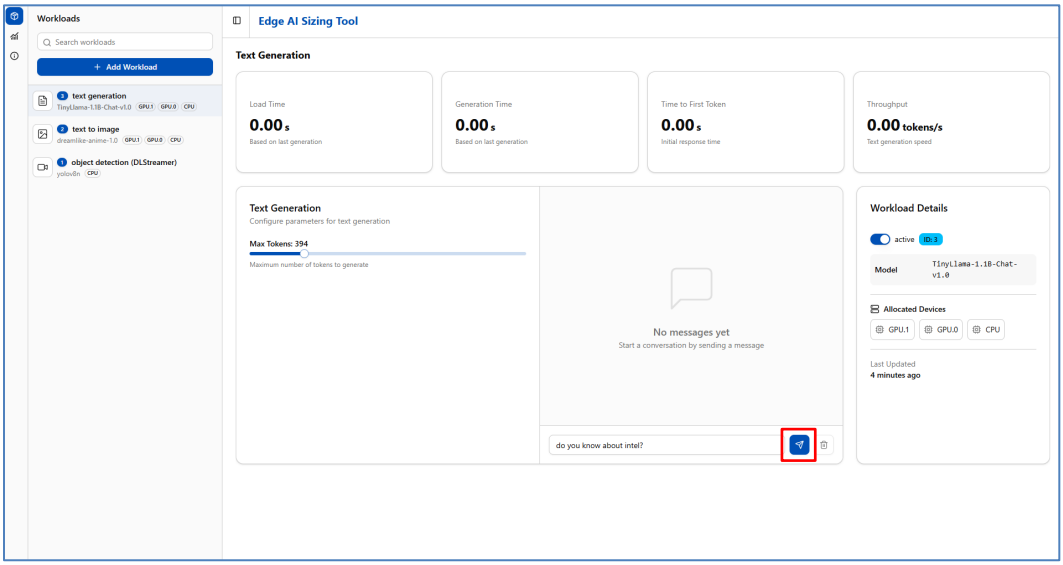
Before executing the inferencing for text generation, the Max Tokens field can be adjusted to specify the maximum number of words to be generated.

Figure 39. Max Tokens Field for Text Generation (NLP)



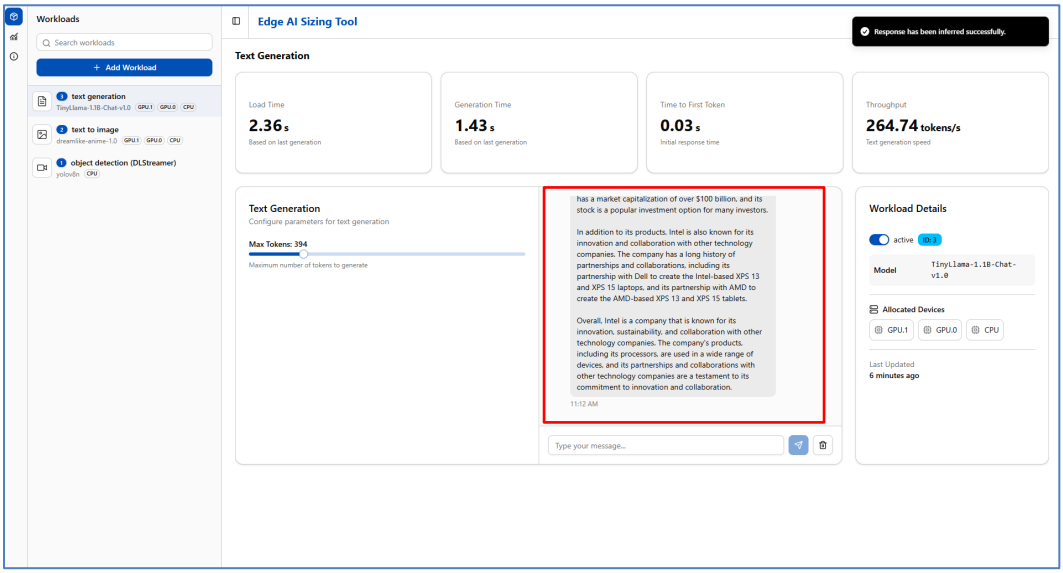
After adjusting the Max Tokens field, click the Send icon button to initiate the text generation inferencing process.

Figure 40. Complete Text Button for Text Generation (NLP)



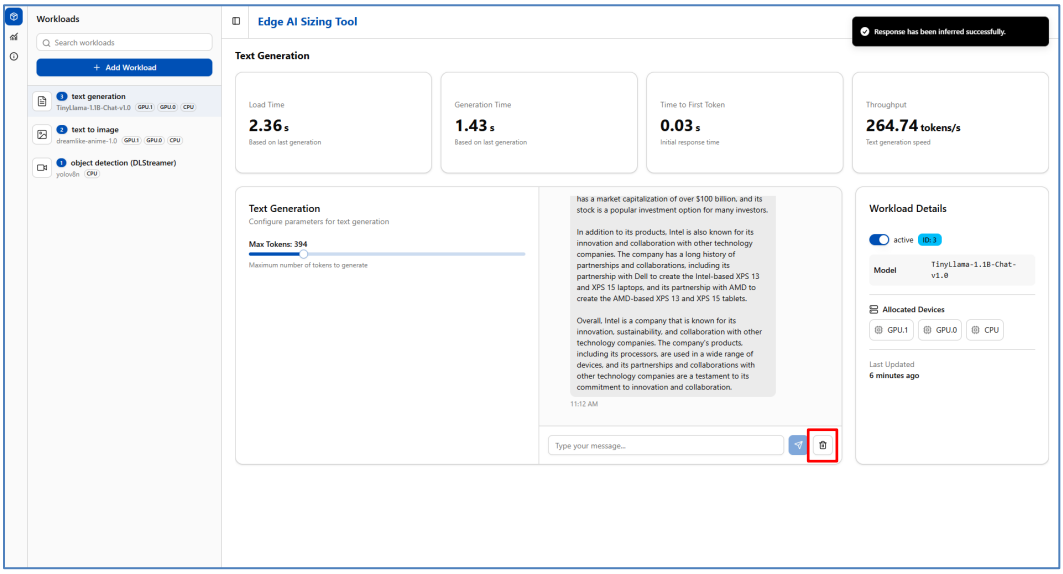
Upon successful completion of the inference, the generated text will be displayed in the text generation placeholder, located right beside the max configuration field and above of the Type Your Message and send icon fields.

Figure 41. Completed Text Result Placeholder for Text Generation (NLP)



To clear the generated text from the inference, click the Dustbin icon button located beside the Send button, where this action will remove all the generated text from the text generation placeholder.

Figure 42. Dustbin Button for Text Generation (NLP)

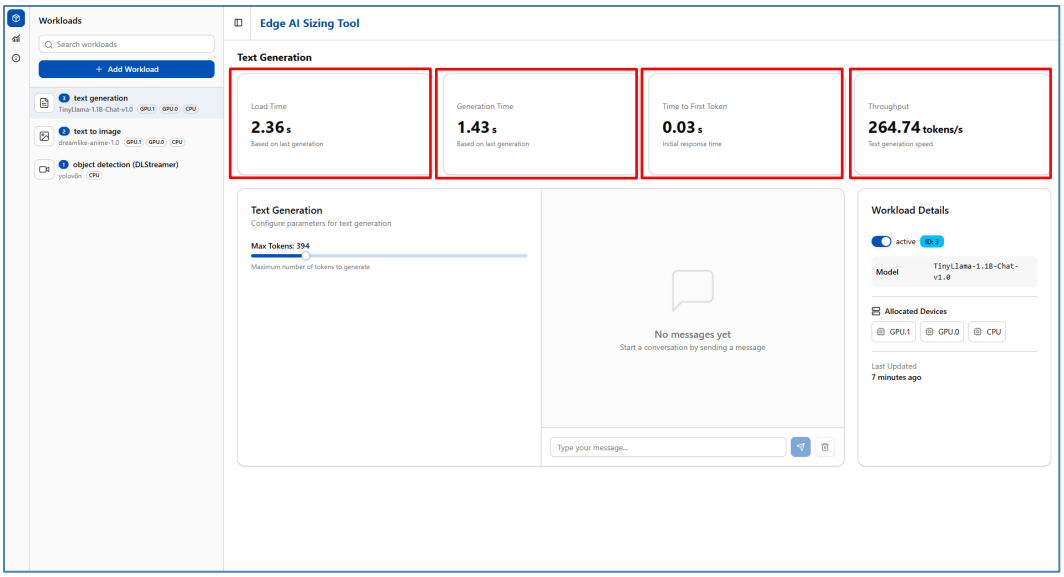


In addition to the generated text, the right panel provides additional information related to the text generation process. This includes:

- **Load Time:** The time taken to load the model onto the accelerator device.
- **Generation Time:** The time taken in seconds to generate the response text.
- **Time to First Token:** The time taken to generate the first token.
- **Throughput:** The rate at which the model can generate tokens per second.

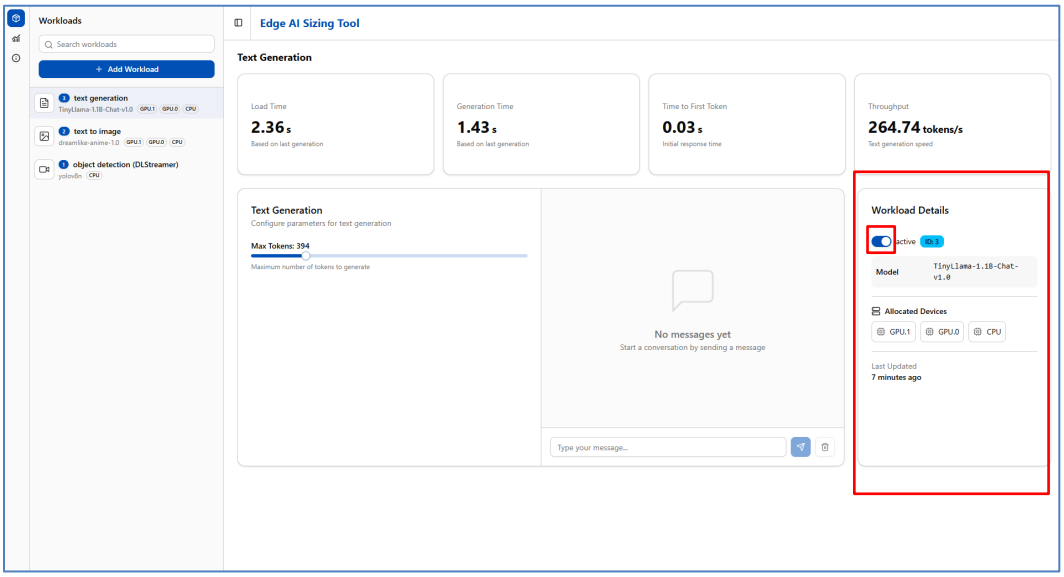
This information can be observed at the top of the right panel.

Figure 43. Load Time, Generation Time, Time to First Token & Throughput for Text Generation (NLP)



In addition to the text generation placeholder, the right panel displays Text Generation workload details, including the workload ID, status (active or inactive), model name, allocated accelerator devices, and the last update timestamp. An active toggle button is available to disable the Text Generation workload that prevents it from being active and running in the system.

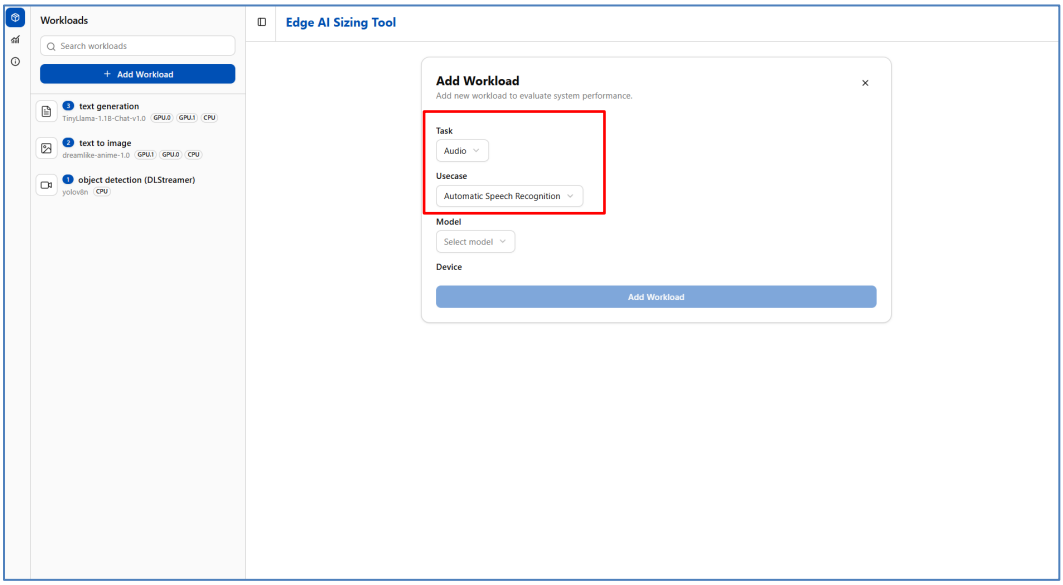
Figure 44. Workload Details for Text Generation (NLP)



5.6 Audio (Automatic Speech Recognition)

To create a workload using the Automatic Speech Recognition use case, you need to select Audio in the Task field and Automatic Speech Recognition in the Use Case field.

Figure 45. Task and Use Case Fields for Text Generation (NLP)



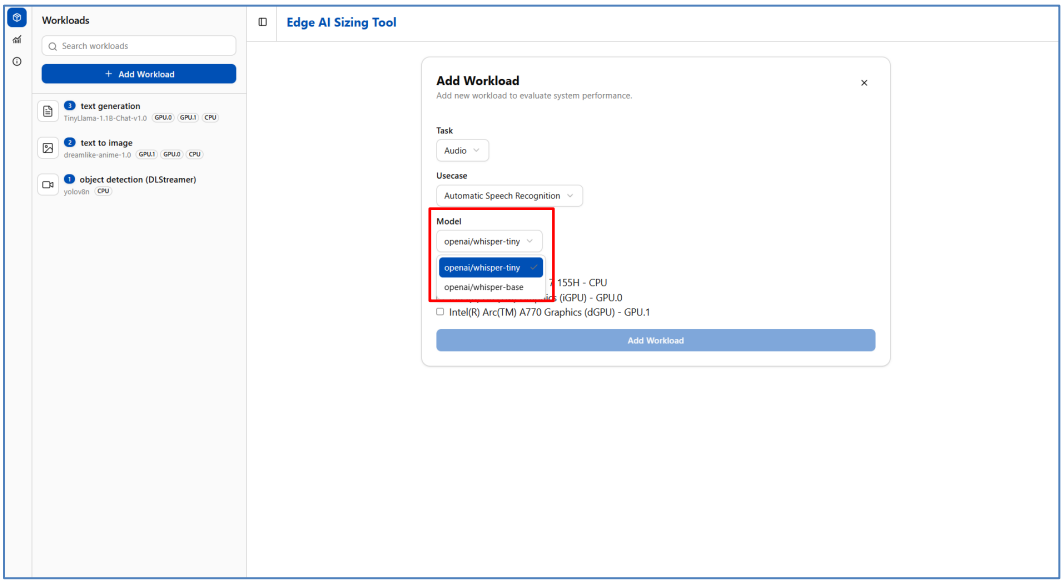
In the Model field, a list of supported and available AI models for Automatic Speech Recognition is provided. One of these models must be selected for speech recognition. The following OpenVINO™ toolkit supported models are available for selection in the Automatic Speech Recognition use case, along with links to where they can be downloaded.

Table 11. Automatic Speech Recognition Supported Models

openai	whisper-tiny	Hugging Face
	whisper-base	Hugging Face

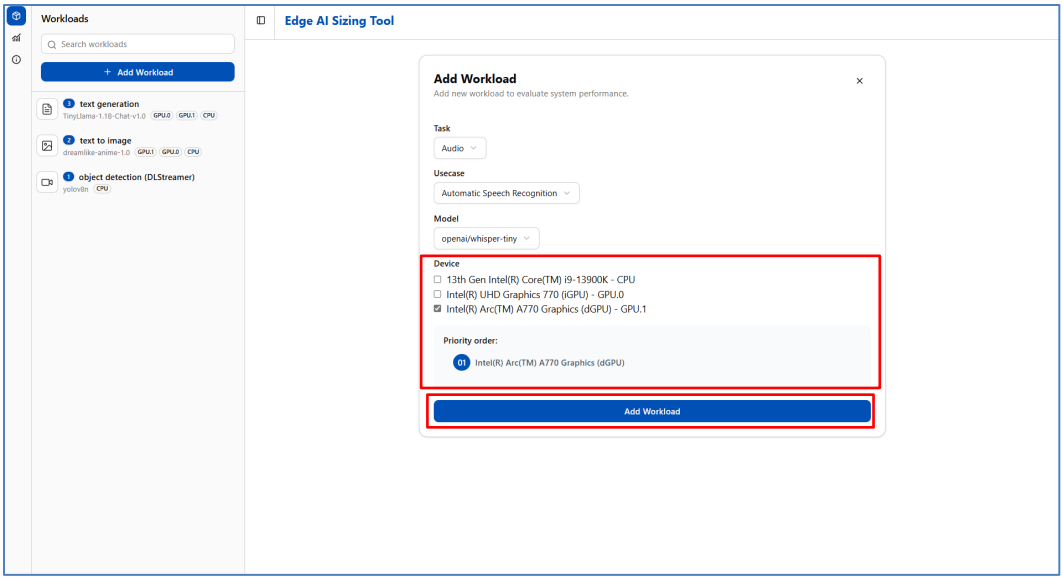
Select the model that best fits the speech recognition requirements, but remember to review the licensing agreements as outlined in the *Disclaimer* under 3.0 Installation.

Figure 46. Model Field for Text Generation (NLP)



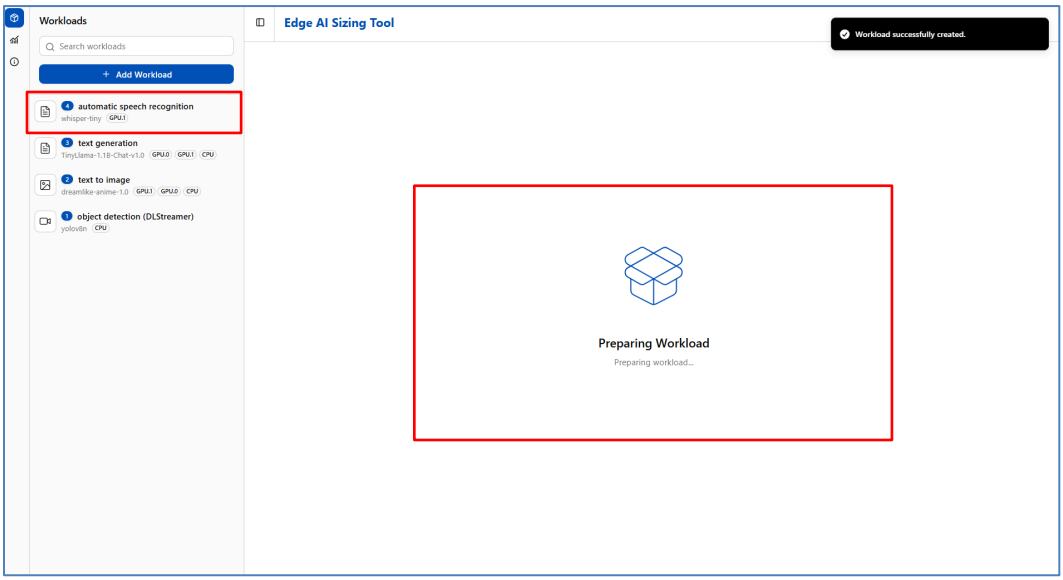
In the Device field, a list of available hardware accelerators is provided, from which a single accelerator must be selected. Once all fields are completed, the Add Workload button will be enabled, and clicking it will save the new workload.

Figure 47. Device Field for Text Generation (NLP)



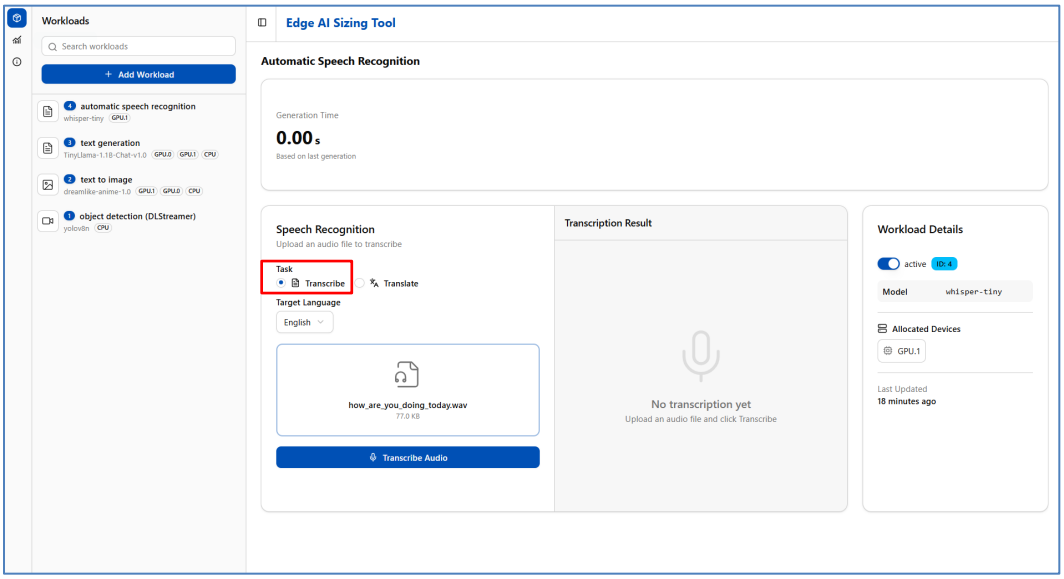
After clicking Add Workload, the right panel will indicate that the tool is preparing the workload. This process may take some time, depending on the speed of the internet connection, as the model will be downloaded in the background if it hasn't been previously downloaded. In the sidebar of the left panel, a new Automatic Speech Recognition workload will be listed, which must be clicked to view the workload content in the right panel.

Figure 48. Preparing Workload Dashboard for Text Generation (NLP)



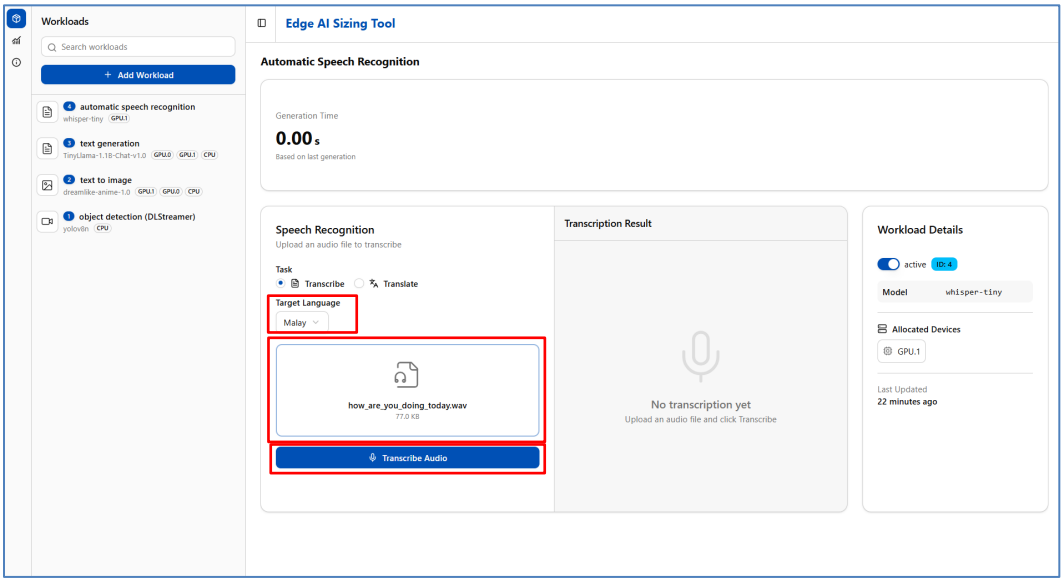
Once the Automatic Speech Recognition workload preparation process is successfully completed, the workload content will be displayed in the right panel. From here, two tasks can be accomplished: Transcribe (transcription translation based on the audio) and Translate (translation to English based on the audio). To perform the Transcribe task, select the Transcribe radio button.

Figure 49. Transcribe Task Button for Text Generation (NLP)



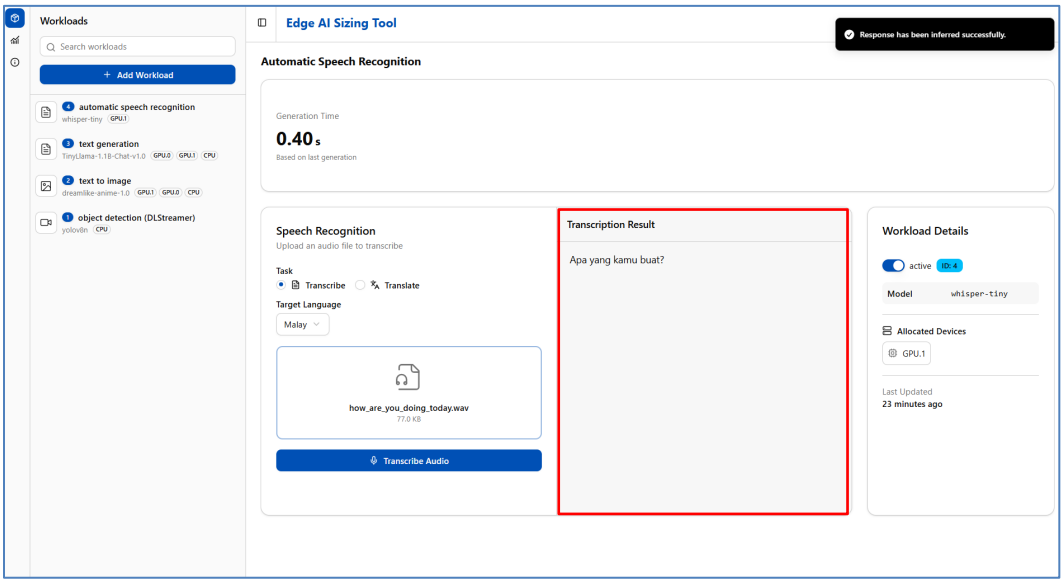
Next, in the Target Language field, select the desired language for the audio transcription. In the audio file field, either use the default audio or upload a file from the system by clicking on the field. Once these selections are made, click the Transcribe Audio button to begin the inferencing process.

Figure 50. Target Language, Audio File Drop Fields & Transcribe Audio Button for Text Generation (NLP)



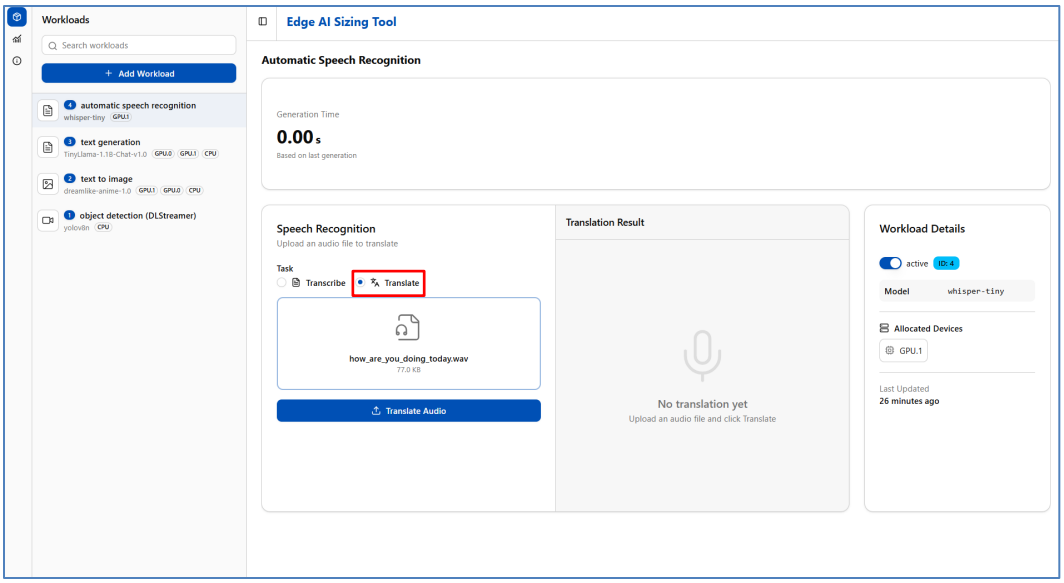
Upon completion of the inferencing for audio transcription, the transcription result will be displayed in the transcription result placeholder next to the configuration form.

Figure 51. Transcription Result Placeholder for Text Generation (NLP)



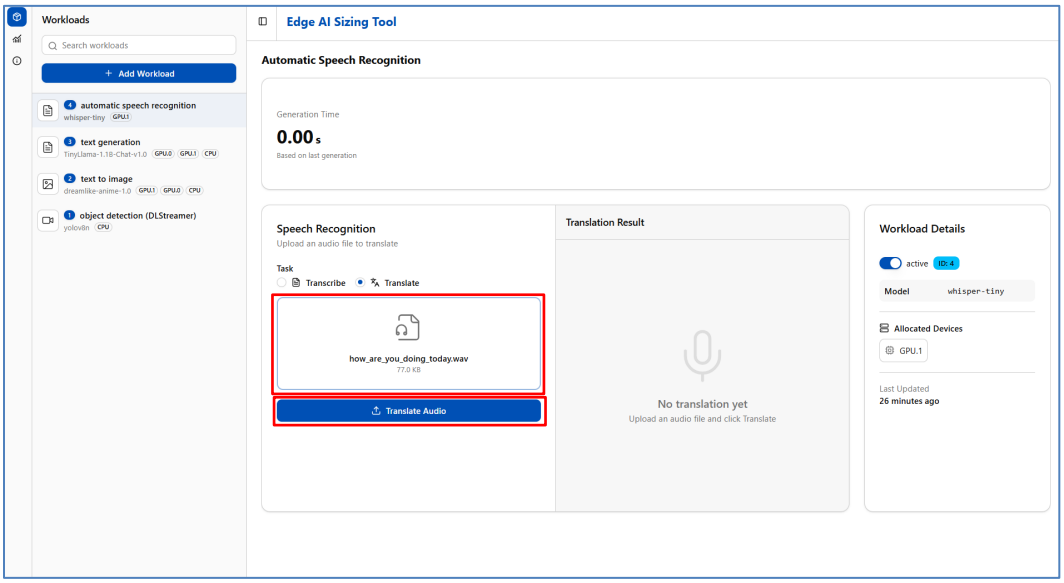
To perform the Translate task, select the Translate radio button.

Figure 52. Translate Task Button for Text Generation (NLP)



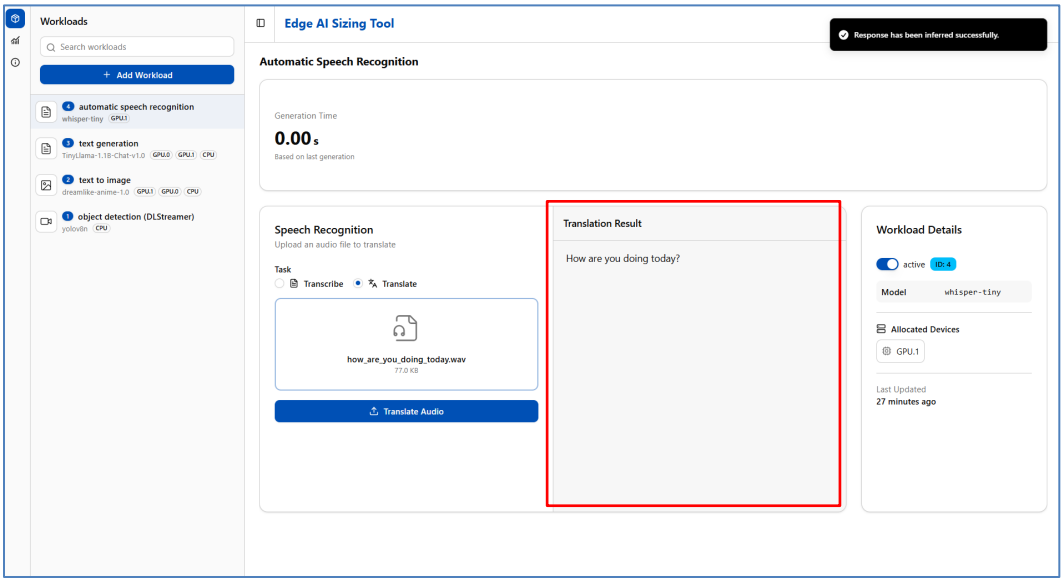
For the audio file field, either use the default audio or upload a file from the system by clicking on the field. Once the audio file is selected, click the Translate Audio button to initiate the inferencing process.

Figure 53. Audio File Drop Field & Translate Audio Button for Text Generation (NLP)



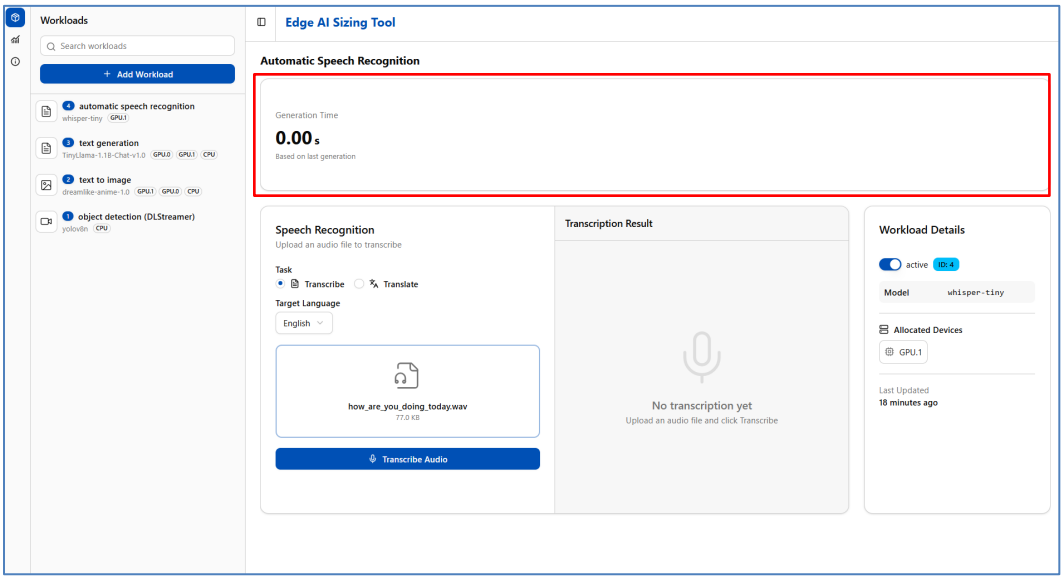
Upon completion of the inferencing for audio translation, the translation result will be displayed in the transcription result placeholder next to the configuration form.

Figure 54. Translation Result Placeholder for Text Generation (NLP)



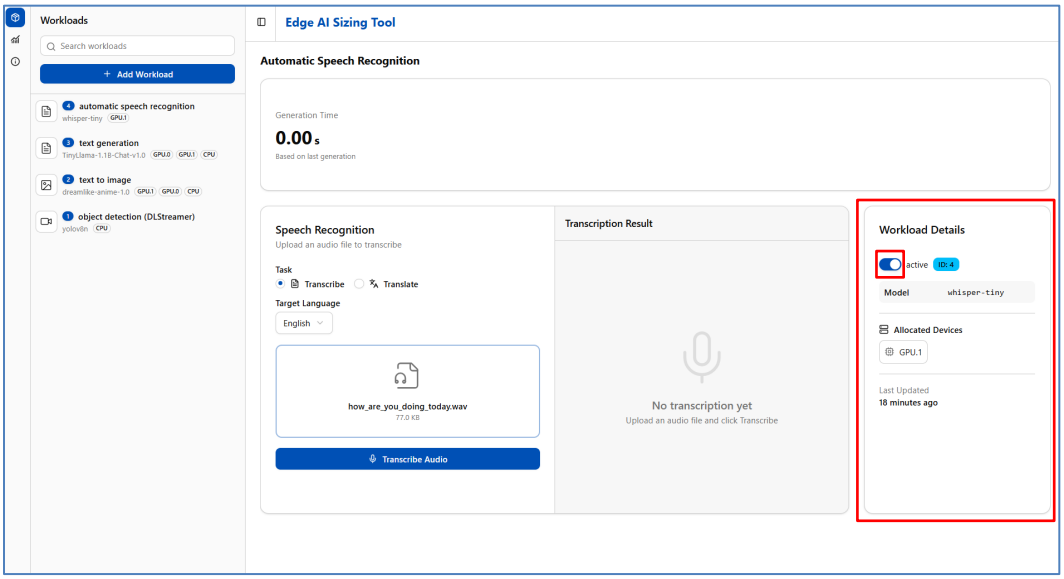
In addition to the audio transcription or translation results, additional information such as generation time is provided. This indicates the time taken in seconds to generate the text and can be observed at the top of the right panel.

Figure 55. Generation Time for Text Generation (NLP)



In addition to the transcription or translation results, the right panel displays details of the Automatic Speech Recognition workload, including the workload ID, status (active or inactive), model name, allocated accelerator devices, and the last update timestamp. An active toggle button is available to disable the workload, preventing it from running in the system.

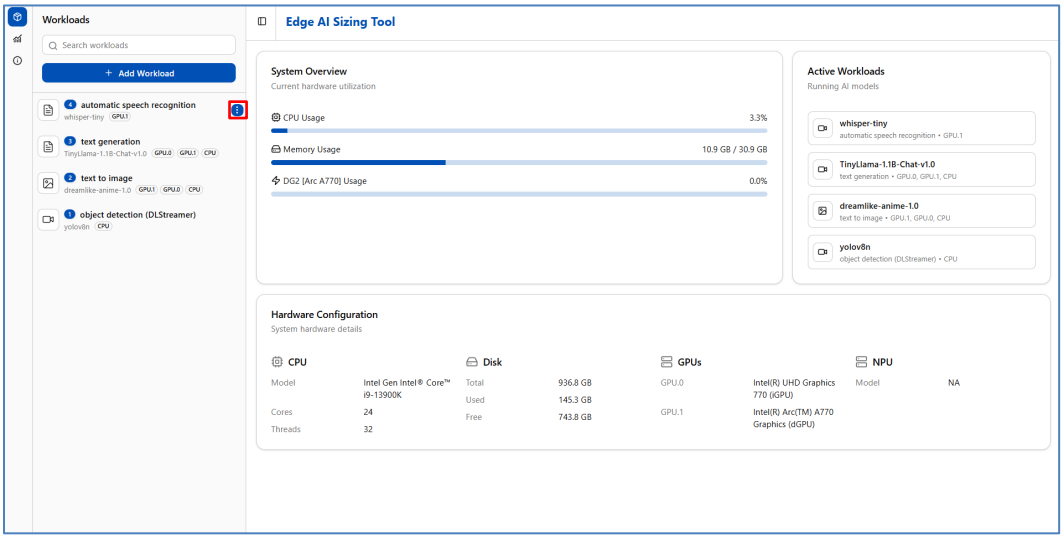
Figure 56. Workload Details for Text Generation (NLP)



5.7 Edit an Existing Workload

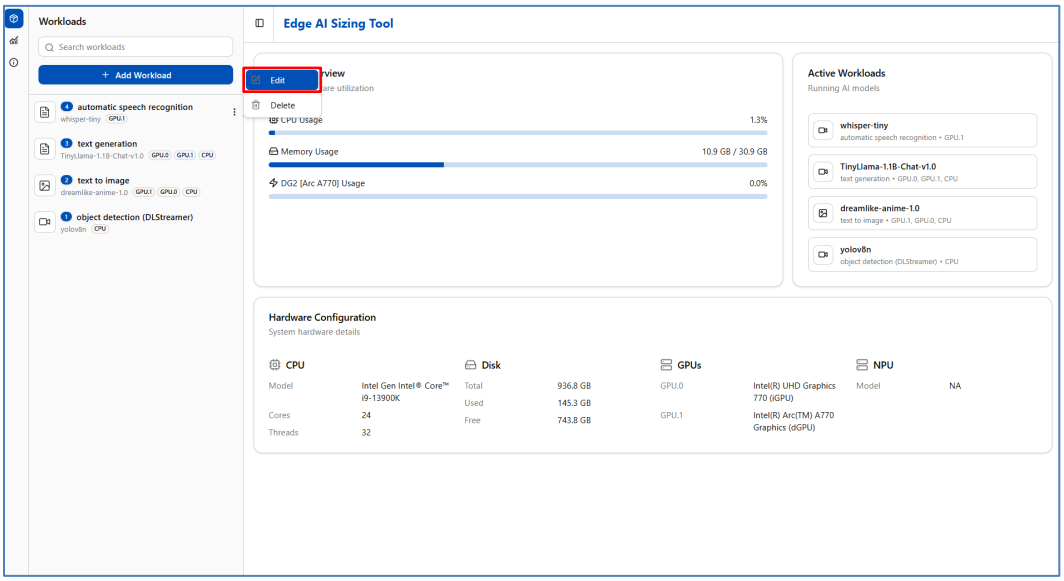
To edit a specific workload, click the Ellipsis icon (three dots) button next to the desired workload in the list on the sidebar under the left panel. This will allow you to access the editing options for that workload.

Figure 57. Ellipsis Button for Workload Edit



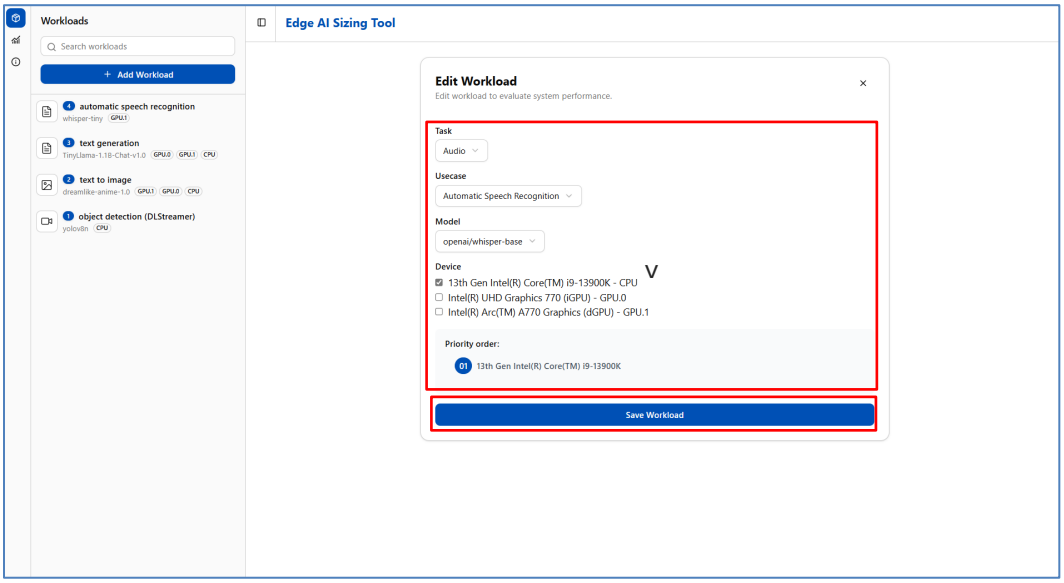
After clicking the Ellipsis icon, a drop-down menu will appear with two options. Select the Edit button to proceed with editing the chosen workload.

Figure 58. Edit Button



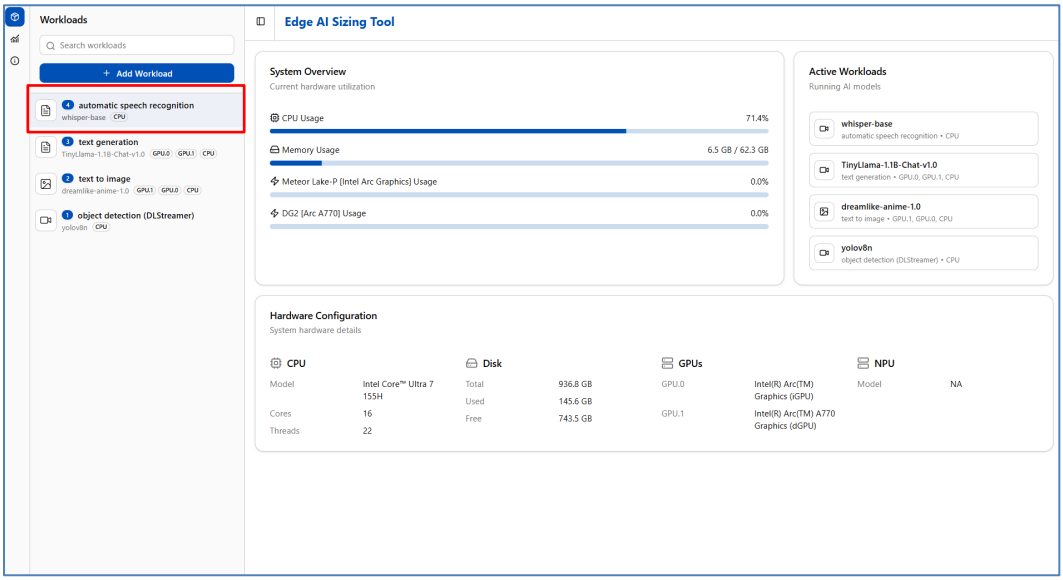
By clicking the "Edit" button, a form will open in the right panel, allowing you to edit the specific workload. All fields in the form must be completed, and once the necessary changes are made, click the "Save Workload" button to update the existing AI workload.

Figure 59. Edit Workload Form



After clicking the Save Workload button, the updated details of the specified workload will be reflected in the sidebar under the left panel, indicating that the changes have been successfully applied.

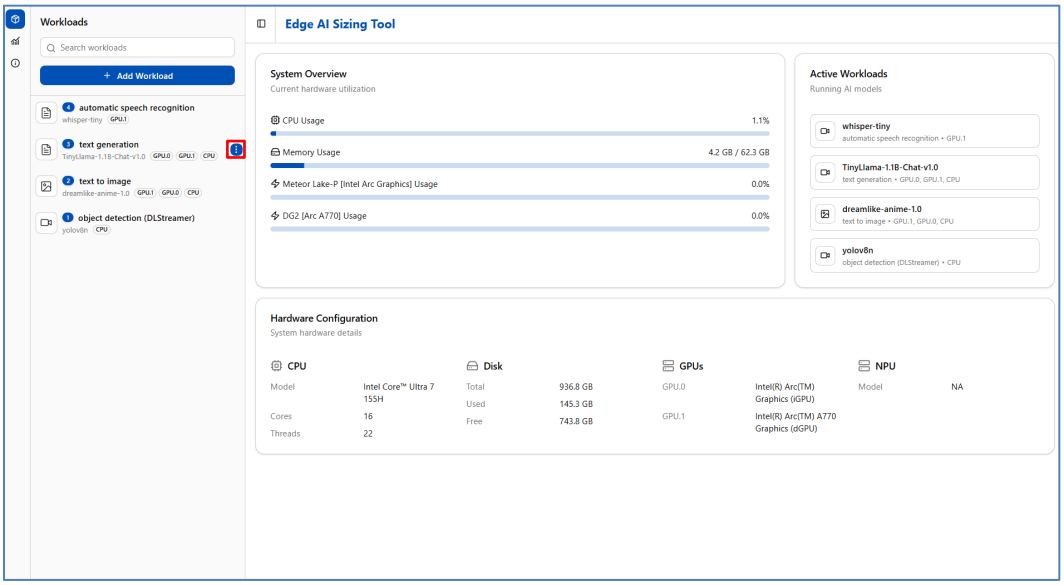
Figure 60. Updated Workload Information



5.8 Delete an Existing Workload

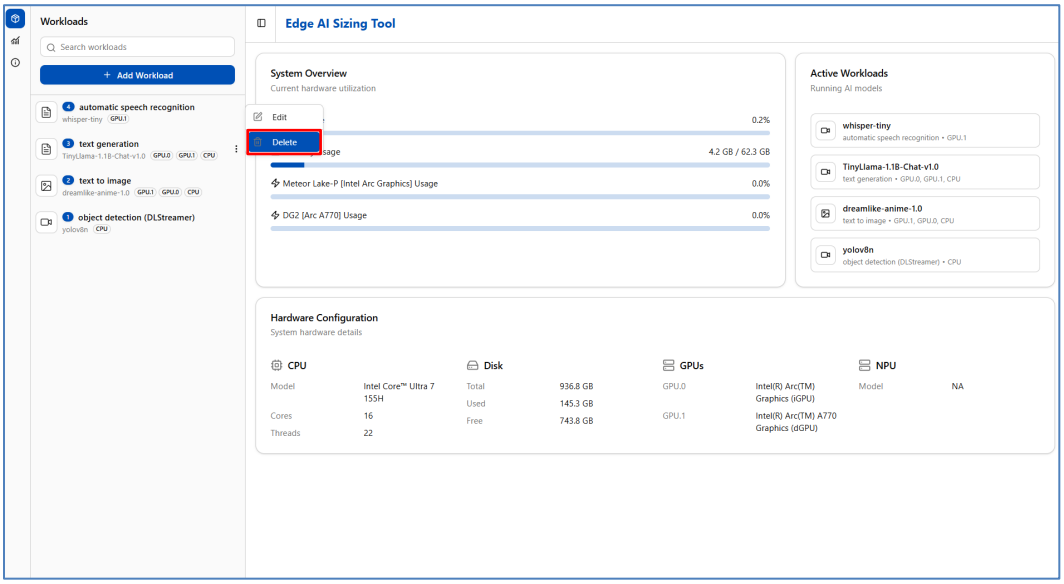
To delete a specific workload, click the Ellipsis icon (three dots) button next to the desired workload in the list on the sidebar under the left panel. This will provide access to the options for removing the workload.

Figure 61. Ellipsis Icon for Workload Deletion



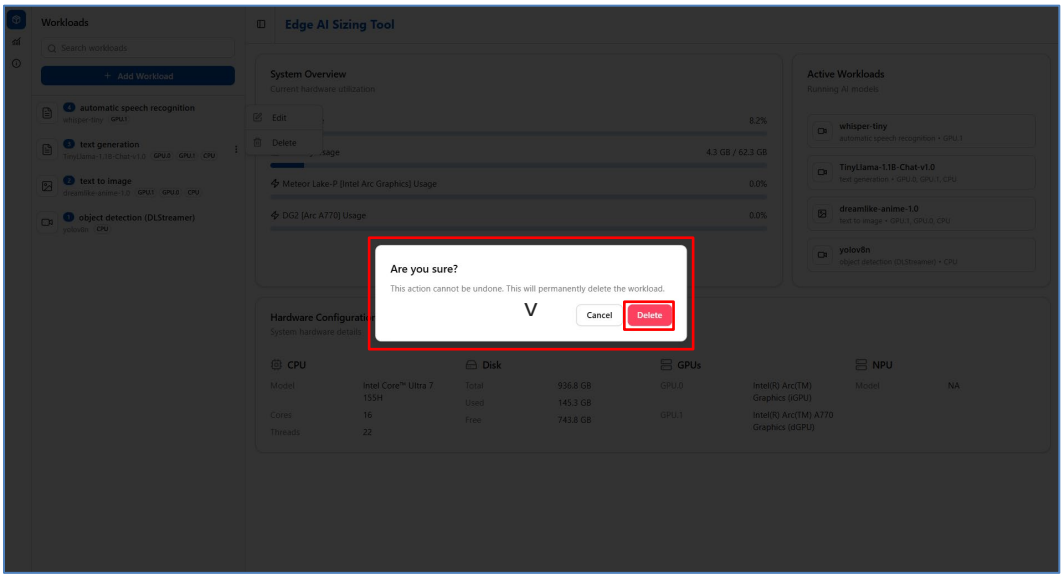
After clicking the Ellipsis icon, a dropdown menu with two options will appear. Select the Delete button to remove the chosen workload.

Figure 62. Delete Button



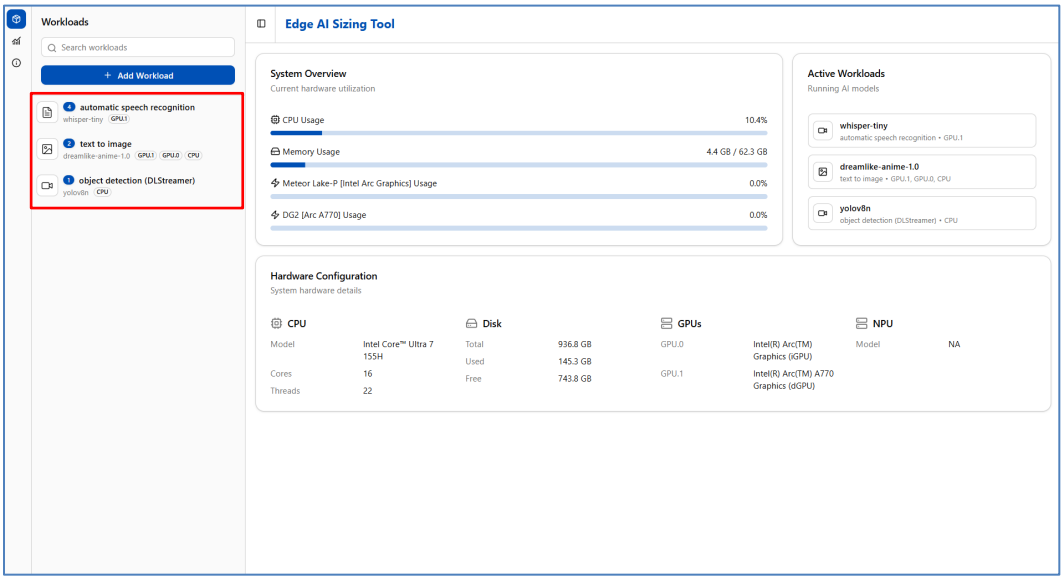
By clicking the Delete button, a pop-up dialog box will appear, prompting you to confirm the action of deleting the workload. Confirm the deletion by clicking the Delete button within the dialog box, and the specific workload will be removed from the tool.

Figure 63. Delete Workload Dialog Box



After confirming the deletion by clicking the Delete button, the specified workload will be removed from the sidebar under the left panel, indicating that it has been successfully deleted from the tool.

Figure 64. Deleted Workload Information



6.0 Known Issues

This section outlines the known issues that have not yet been resolved or addressed with solutions.

6.1 Limitations

This application has been validated and tested on the hardware listed in the documentation. While we strive to ensure compatibility and performance, the application may not function as expected on other hardware configurations. Users may encounter issues or unexpected behavior when running the application on untested hardware. If you encounter any issues or have suggestions for improvements, you can create an issue on an open-source GitHub repository. Some of the already known issues/limitations are:

1) The intel-gpu-tools package does not support Intel® Arc™ B-Series Graphics Cards, resulting in the inability to display GPU utilization metrics.

2) iGPU utilization and device name may not be able to be shown on Intel® Core™ Ultra 9 288V.

3) Object Detection Inferencing Windows may not show results after some period of hours running.

4) System Overview and System Monitor may only show the PCI ID (for example, e20b) for certain GPU models instead of the actual descriptive name.

5) Text generation may produce gibberish and illogical output for some LLM models when using Intel® Arc™ Ultra 7 Processor 155H iGPUs.