MAI-Image-2 Model Card

***Date**: March 18, 2026*

Model summary

Developer	Microsoft Ireland Operations Limited (MIOL) 70 Sir John Rogerson's Quay, Dublin 2, D02 R296, Ireland
Description	The model has a diffusion-based generative architecture designed for text-to-image synthesis. It operates by progressively transforming random noise into a coherent image that aligns with a given text prompt. This approach leverages a flow-matching loss to learn a continuous transformation between the noise distribution and the data distribution, ensuring stable and efficient training. This combination of flow-matching objectives and diffusion inference enables the model to produce high-quality, diverse images that maintain strong alignment with the input text, making it suitable for creative generation, design tasks, and multimodal applications.
Model architecture	The model has a diffusion-based generative architecture for text-to-image synthesis.
Parameters	10B-50B non-embedding parameters
Inputs	Text input
Context length	32K tokens
Outputs	Image output. Maximum 1024x1024 pixels.
Public data summary (or summaries)	Public data summary available here
Training Dates	January 2026 to March 2026
Release date Release date in the EU (if different)	MAI Playground – 19 March 2026
License	Various product and service terms where the model is deployed, such as those for MAI Playground.

Model dependencies:	N/A
List and link to any additional related assets	N/A
Acceptable use policy	As described in the License section above. Subsequent integrations of MAI-Image-2 may be subject to different terms of service and policies.

Model overview

The model has a diffusion-based generative architecture designed for text-to-image synthesis. It operates by progressively transforming random noise into a coherent image that aligns with a given text prompt. This approach leverages a flow-matching loss to learn a continuous transformation between the noise distribution and the data distribution, ensuring stable and efficient training. This combination of flow-matching objectives and diffusion inference enables the model to produce high-quality, diverse images that maintain strong alignment with the input text, making it suitable for creative generation and design tasks.

Alignment approach

Our alignment objective for MAI-Image-2 is to reduce the generation of harmful or inappropriate (e.g., violent, gory, sexual) images, even when requested by a user. We took a defense-in-depth approach, applying mitigations to the data during model development, and deploying the model with additional safety mitigations. The initial release of MAI-Image-2 is through integrations with Microsoft products and services, such as MAI Playground.

Usage

Primary use cases and out-of-scope uses

MAI-Image-2 is a general-purpose text-to-image generative model, intended for creative generation and design tasks. The model is particularly capable at generating photorealistic imagery.

The initial deployment of MAI-Image-2 is through integrations with Microsoft products and services, and those products and services should be used according to the relevant terms. For example, the initial integration of MAI-Image-2 is in MAI Playground and the relevant terms of service can be found here. In addition to the terms of service, Microsoft AI products and services should be used in accordance with any relevant codes of conduct.

Distribution channels

The model will initially be made available for use by end users as part of Microsoft AI products and services, such as MAI Playground, a publicly available site for users to interact with MAI models.

Any future release formats, such as an API release, will be accompanied by an update to relevant documentation.

Responsible AI considerations

Despite technical mitigations such as data filtering, image generation models are known to produce harmful or unexpected content based on user requests. In addition to technical work on the model such as data filtering, additional mitigations are also applied at the system level to further enhance end user safety (e.g., content classifiers). Some common risk areas associated with image generation models include violent or gory content, sexual content or nudity, depictions of public figures, replication of trademarked or other protected material. Evaluations of specific content areas are described below and are observed as similar to comparable image generation models.

Data overview

Training, testing, and validation datasets

The training dataset consists of paired images and text descriptions, where each caption provides a detailed account of the visual content. The data spans a broad, general-purpose domain rather than being restricted to a specialized field like healthcare or finance. It includes everyday objects, natural scenes, people, and abstract concepts, making it suitable for open-domain text-to-image generation. In terms of modality, the dataset combines visual data (images) with natural language text, enabling multimodal learning. The text descriptions are typically concise yet descriptive, capturing key attributes such as objects, actions, and scene context. These characteristics—broad domain coverage and multimodal pairing—are directly aligned with the model's purpose: to generate high-quality, semantically aligned images from text prompts across a wide

range of scenarios. To read more about the data used to train MAI-Image-2 please see the public data summary.

Quality and performance evaluation

The model was evaluated by human raters alongside comparable models and across a range of capability areas. Specifically, raters were tasked with selecting a preferred model output, in different topic areas based on real user intents (for example, "product/branding," "cartoon," "photorealistic") and by reference to the output's alignment with the prompt intent as well as the output's visual appeal. This resulted in an Elo score calculation. Generally, the model was found to perform as well or to exceed the performance of the comparable models, performing particularly well when generating photorealistic imagery.

Category	MAI-Image-2	MAI-Image-1
Photorealistic & Cinematic Imagery	1201 ± 12	1104 ± 5
Product, Branding, commercial design	1191 ± 11	1085 ± 5
3D Imaging & Modeling	1184 ± 22	1096 ± 8
Cartoon, Anime & Fantasy	1186 ± 14	1100 ± 5
Art	1191 ± 18	1104 ± 7
Portraits	1201 ± 17	1095 ± 6
Text rendering	1186 ± 12	1069 ± 5
Overall	1190 ± 8	1093 ± 4

Safety evaluation and red-teaming

Image generation models are known to produce potentially harmful or unexpected content based on user requests, with common risk areas including violent or gory content, sexual content or nudity. In addition to technical work on the model such as data filtering, we evaluated the model focusing on safeguards in place in the product deployments.

The Microsoft AI Red Team (AIRT) conducted multiple rounds of red teaming of MAI-Image-2 to emulate real world adversaries across a spectrum of skill levels, ranging from straight forward prompting to advanced attack methodologies. AIRT developed attack strategies spanning eight harm categories and engaged subject matter experts to evaluate the model against these risk areas.

Evaluation was carried out in two phases: pre-mitigation and post-mitigation. The assessment followed a break-fix cycle in which AIRT identified vulnerabilities and shared them with the model development team for remediation. The updated model was then re-evaluated in subsequent rounds to assess the effectiveness of mitigations and to identify any remaining weaknesses.

Following multiple rounds of mitigation and re-testing, the number of findings decreased significantly. Residual findings generally required moderate to high levels of adversarial effort to reproduce.