



Categories



Public Summary of Training Content

Version of Summary: [Version #1](#)

Last update: [March 17, 2026](#)

1. General information

1.1. Provider identification

Provider name and contact details: [Midjourney, Inc.](#)

Authorised representative name and contact details: [Mark Foster \(mark.foster@samadvisory.eu\)](#)

1.2. Model identification

Versioned model name(s): [Midjourney Image and Video family of models](#)

Model dependencies: [Midjourney V8, V8.1, and V8.2](#)

Date of placement of the model on the Union market: [March 17, 2026](#)

1.3. Modalities, overall training data size and other characteristics

Modality	Training data size	Types of content
☒ Text	☒ Less than 1 billion tokens ☐ 1 billion to 10 trillions	The family of models datasets containing pu

Ask AI



	tokens ☐ More than 10 trillions tokens	available text annotations and image captions.
☒ Image	☐ Less than 1 million images ☐ 1 million to 1 billion images ☒ More than 1 billion images	The family of models trains on datasets containing photography, visual art works, illustrations, textual metadata associated with images, human-provided annotations, prompts and preference data.
☐ Audio	☐ Less than 10 000 hours ☐ 10 000 to 1 million hours ☐ More than 1 million hours	
☒ Video	☐ Less than 10 000 hours ☐ 10 000 to 1 million hours ☒ More than 1 million hours	The family of models trains on datasets containing video clips, video effects, textual metadata associated with these videos, human-provided annotations, prompts and preference data.
☐ Other		

Latest date of data acquisition/
collection for model training:

The data used to train the model includes datasets with varying cutoff dates. Datasets were used to train the models as late as March 2026. The model is continuously trained and may undergo additional fine-tuning which may be released in new versions.

Description of the linguistic
characteristics of the overall
training data:

Training sources include both European and non-European languages.

Other relevant characteristics
of the overall training data:

Midjourney training data represents a large-scale and diverse range of data including publicly-available websites, images, text, and video. The datasets are fine tuned for Midjourney's purposes. Midjourney training data undergoes several processing steps during training, including: deduplication, removal of low quality images, safety filtering, privacy processing to filter or remove sensitive personal information, and categorization by relevance, quality, or image format.

Ask AI



Additional comments
(optional):

N/A

2. List of data sources

2.1. Publicly available datasets

Have you used publicly
available datasets to train the
model?

☒ Yes ☐ No

If yes, specify the modality(ies)
of the content covered by the
datasets concerned:

☒ Text ☒ Image ☐ Video ☐ Audio
☐ Other *If so, please specify...*

List of large publicly available
datasets:

Midjourney datasets are composed of a wide
variety of data publicly accessible online.

General description of other
publicly available datasets not
listed above:

Public datasets are filtered and fine tuned for
quality and safety as described above and
exclude sources that have opted out of training
using web controls such as robots.txt files.

Additional comments
(optional):

N/A

2.2. Private non-publicly available datasets obtained from third parties

2.2.1. Datasets commercially licensed by rightsholders or their representatives

Have you concluded
transactional commercial
licensing agreement(s) with
rightsholder(s) or with their
representatives?

☒ Yes ☐ No

If yes, specify the modality(ies)
of the content covered by the
datasets concerned:

☒ Text ☒ Image ☐ Video

2.2.2. Private datasets obtained from other third parties

Have you obtained private

☒ Yes ☐ No

Ask AI



datasets from third parties that are not licensed as described in Section 2.2.1, such as data obtained from providers of private databases, or data intermediaries?

If yes, specify the modality(ies) of the content covered by the datasets concerned:

☒ Text ☒ Image ☐ Video ☐ Audio
☐ Other *If so, please specify...*

If publicly known, list private datasets obtained from other third parties:

Some Midjourney datasets are purchased or licensed from third parties. These deals are bound by confidentiality obligations.

General description of non-publicly known private datasets obtained from third parties

Data is covered by agreements that outline the party's roles and responsibilities with respect to the datasets Midjourney uses.

Additional comments (optional):

N/A

2.3. Data crawled and scraped from online sources

Were crawlers used by the provider or on behalf of?

☒ Yes ☐ No

If yes, specify crawler name(s)/ identifier(s):

Deals are bound by confidentiality obligations

Purposes of the crawler(s):

Crawlers are used to obtain publicly available content for training purposes.

General description of crawler behaviour:

Crawlers are used to find information, scan websites, and extract data from lawfully and publicly accessible resources. Crawlers are designed to respect robots.txt rules and do not circumvent technological measures.

Period of data collection:

2023 to present

Comprehensive description of the type of content and online sources crawled:

Crawled data pulls from publicly available online material including images and associated text descriptions. The crawled data was filtered for safety, quality, and privacy.

Type of modality covered:

☒ Text ☒ Image ☒ Video ☐ Audio
☐ Other *If so, please specify..*

Ask AI



Summary of the most relevant domain names crawled:

The crawled data includes data from publicly available online websites and sources that encompass a large variety of content types and languages. Toplevel domain names crawled include: .com, .org, and .net as well as other global sites.

Additional comments (optional):

N/A

2.4. User data

Was data from user interactions with the AI model (e.g. user input and prompts) used to train the model?

☒ Yes ☐ No

Was data collected from user interactions with the provider's other services or products used to train the model?

☐ Yes ☒ No

If yes, provide a general description of the provider's services or products that were used to collect the user data:

N/A

Type of modality covered:

☒ Text ☒ Image ☐ Video ☐ Audio
☐ Other *If so, please specify...*

Additional comments (optional):

Midjourney adheres to its [Privacy Policy](#) and [Terms of Service](#), as applicable, for user data training.

2.5. Synthetic data

Was synthetic AI-generated data created by the provider or on their behalf to train the model?

☒ Yes ☐ No

If yes, modality of the synthetic data:

☒ Text ☒ Image ☐ Video ☐ Audio
☐ Other *If so, please specify...*

Ask AI



If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:

Midjourney may use internal models to generate synthetic data for training.

Information about other AI models, including provider's own AI model(s) not available on the market, used to generate synthetic data to train the model to which this Summary applies:

N/A

Additional comments (optional):

N/A

2.6. Other sources of data

Have data sources other than those described in Sections 2.1 to 2.5 been used to train the model?

☒ Yes ☐ No

If yes, provide a narrative description of these data sources and the data:

Midjourney used datasets that it has acquired through its business operations.

Additional comments (optional):

N/A

3. Data processing aspects

3.1. Respect of reservation of rights from text and data mining exception or limitation

Are you a Signatory to the Code of Practice for general-purpose AI models that includes commitments to respect reservations of rights from the TDM exception or limitation?

☐ Yes ☒ No

Ask AI



Describe the measures implemented before model training to respect reservations of rights from the TDM exception or limitation before and during data collection, including the opt-out protocols and solutions honoured by the provider or, as applicable, by third parties from which datasets have been obtained:

Midjourney implemented safety and quality filtering, content moderation, honored robots.txt files instructions where appropriate, and conducted privacy processing to filter or remove sensitive personal information from training data.

Additional comments (optional):

N/A

3.2. Removal of illegal content

General description of measures taken:

Midjourney adheres to applicable laws and best practices with respect to removing illegal content. Midjourney training data undergoes safety filtering to remove certain data with known risk of containing child sexual abuse material (CSAM) and other categories of sensitive or disallowed content. See also rules of conduct for its users at: <https://>

Midjourney Website

Midjourney Discord Server

3.3. Other information (optional)

Other relevant information about data processing (optional):

N/A

Ask AI

