Public Summary of Training Data Content for Seedance 2.0

Version of the Summary:	1.0
Last update:	31 July 2026

1. General information

1.1 Provider identification

Provider name and contact details:	ByteDance Nexus AI Pte. Ltd (**ByteDance**)
Authorised representative name and contact details:	Mikros Information Technology Ireland Limited

1.2 Model identification

Versioned model name(s):	Seedance 2.0
Model dependencies:	N/A
Date of placement of the model on the Union market:	March 2026

1.3 Modalities, overall training data size and other characteristics

Modality	Training data size	Types of content
Text	1 billion to 10 trillion tokens	

		The model was pre-trained using a large-scale, diverse collection of data spanning a wide-range of domains and modalities, including textual captions.
Image	More than 1 billion images	The model was pre-trained using a large-scale, diverse collection of data spanning a wide-range of domains and modalities, including infographics and imagery.
Audio	More than 1 million hours	Most audio content originates from the associated video files, and only a limited subset consists of standalone audio.
Video	More than 1 million hours	Seedance 2.0's training dataset comprised video with accordingly textual description.
Other	N/A	N/A

Latest date of data acquisition/collection for model training:	February 2026
Description of the linguistic characteristics of the overall training data:	No specific geographic region was intentionally excluded from the data collection process. Video data contains a diverse range of global languages, including languages from within Europe.
Other relevant characteristics of the overall training data:	Seedance 2.0's training dataset comprised clips of video with accordingly textual description. For model training purposes, ByteDance uses a large-scale mixture of publicly available data, licensed data, and synthetic data, designed to ensure broad coverage of domains and contexts. The dataset has been curated with the objective of maximizing representativeness and robustness, while applying filtering mechanisms to remove low-quality or harmful content where appropriate.
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
If yes, specify the modality(ies) of the content covered by the datasets concerned:	Image, video, audio
List of large publicly available datasets:	The training data corpus comprises a mix of publicly available and licensed data. This may include image, video and audio datasets made available by third parties through public repositories and online platforms. As described above, such datasets are curated and pre-processed with the objective of maximizing representativeness and robustness, while applying filtering mechanisms to remove low-quality or harmful content where appropriate.
General description of other publicly available datasets not listed above:	See the description of linguistic characteristics and other relevant characteristics of overall training data in Section 1.
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
If yes, specify the modality(ies) of the content covered by the datasets concerned:	Image, video and audio

2.2.2 Private datasets obtained from other third parties

Have you obtained private 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?	Yes
If yes, specify the modality(ies) of the content covered by the datasets concerned:	Image, video and audio
If publicly known, list private datasets obtained from other third parties:	N/A
General description of non-publicly known private datasets obtained from third parties	We have obtained datasets from third-parties on a licensed basis. Such datasets comprise content across image, video and audio modalities.
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
If yes, specify crawler name(s)/identifier(s):	Bytespider
Purposes of the crawler(s):	In the context of ByteDance's AI model development, Bytespider is used to collect publicly accessible image, video and audio data to support training, testing, and validation of the AI model.
General description of crawler behaviour:	ByteDance's crawler is configured to avoid overloading websites and to operate in a manner consistent with ethical web scraping practices. Our crawler is designed to respect robots.txt instructions, and not to circumvent control measures such as paywalls or to access password-protected content.
Period of data collection:	2025
Comprehensive description of the type of content and online sources crawled:	As above, we may use our crawler to collect publicly accessible data. The crawled content primarily consists of images and video.
Type of modality covered:	Image and video

Summary of the most relevant domain names crawled:	The most relevant domains crawled includes resources spanning natural scenes, objects, people and diagrams.
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?	No
Was data collected from user interactions with the provider’ s other services or products used to train the model?	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:	N/A
Additional comments (optional):	N/A

2.5 Synthetic data

Was synthetic AI-generated data created by the provider or on their behalf to train the model?	Yes
If yes, modality of the synthetic data:	Text (description of videos and images)
If yes, specify the general-purpose AI model(s) used to generate the synthetic data if available on the market:	Synthetic data was generated using a range of general-purpose AI models, including vision language models and large language models.
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:	We may use fine-tuned versions of internal models to generate synthetic data.
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?	No
If yes, provide a narrative description of these data sources and the data:	N/A
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?	No
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:	ByteDance's crawler is configured to avoid overloading websites and to operate in a manner consistent with ethical web scraping practices. Our crawler is designed to respect robots.txt instructions, and not to circumvent control measures such as paywalls or to access password-protected content.
Additional comments (optional):	N/A

3.2 Removal of illegal content

General description of measures taken:	As mentioned, we applied preprocessing and filtering methods, including filtering models, to exclude unsafe or harmful content.

3.3 Other information (optional)

Other relevant information about data processing (optional):	N/A