A More Results

In this section, we present experimental results in table format. The overall performance in MM-NIAH is shown in Tab. 2, which is obtained by averaging the performance across the six tasks in MM-NIAH. We also provide the performance of each task in Tab. 6 to Tab. 11. The performance for each context length range is obtained by averaging the accuracy of that context length range across different needle depths. For samples containing multiple needles, we average the depths of each needle to serve as the needle depth of this sample.

Table 2: **Overall performance on MM-NIAH for each context length range.**

Model	**1K**	**2K**	**4K**	**8K**	**12K**	**16K**	**24K**	**32K**	**40K**	**48K**	**64K**	**Overall**
Emu2-Chat	33.0	27.8	17.2	5.9	0.9	0.0	0.0	0.0	0.0	0.0	0.0	7.7
VILA-13B	44.7	39.3	34.9	28.3	22.0	8.9	1.1	0.2	0.1	0.0	0.1	16.3
IDEFICS2	48.0	33.8	16.4	13.8	14.3	1.2	0.0	0.0	0.0	0.0	0.0	11.6
LLaVA-1.6-13B	47.0	45.0	41.6	35.0	24.3	15.5	5.7	0.8	0.2	0.1	0.0	19.6
LLaVA-1.6-34B	57.9	53.5	47.1	38.6	27.0	8.2	0.0	0.0	0.0	0.0	0.0	21.1
InternVL-1.5	59.5	55.3	50.1	46.4	45.2	41.9	39.5	33.2	31.6	33.2	30.1	42.4
InternVL-1.5-RAG	67.5	61.1	53.3	51.2	50.6	51.5	46.2	46.2	43.8	40.1	39.0	50.1
Gemini-1.5	64.7	58.3	56.8	57.1	55.4	53.7	53.6	51.9	52.5	50.7	53.6	55.3
GPT-4V	-	-	-	-	-	-	-	-	-	-	-	-
Human	99.7	99.1	97.9	99.0	98.5	98.8	99.9	99.4	99.2	98.6	98.5	98.9

A.1 More findings

In addition to the findings discussed in Section 4.2, we provide more findings here.

Placing questions before context does NOT improve model performance. As shown in Fig. 3, all models perform poorly in understanding image needles, which we attribute to the fact that models struggle to remember the details of each image in a long multimodal document. An intuitive improvement method is placing the question before the context, which allows the model to see the options first and then read the document. However, as illustrated by the error cases (see the first row in Fig. 5), this approach cause models like InternVL-1.5 to fail in following the instructions in the questions. In fact, we observe that this phenomenon holds for all MLLMs, resulting in near-zero performance. Therefore, we do not provide quantitative results but qualitatively analyzed this issue.

The long context understanding ability of Gemini-1.5 is not perfect. As claimed by Gemini-1.5 [35], their model achieves near-perfect performance in long context evaluation with video haystack. However, in our benchmark, their model still performed poorly. Notably, in the video haystack, they only insert textual information by overlaying the text “The secret word is "needle"” on a single randomly sampled video frame in a 10.5 hour video. In contrast, in our benchmark, we inserted another image as additional visual information into the images. Furthermore, their video haystack consists of long videos, whereas our multimodal haystack comprises long multimodal documents. These differences both contribute to the performance decline of Gemini-1.5 in our benchmark.

Multimodal fine-tuning without long context data impairs model’s ability to handle long context. We provide the comparison of InternLM2-20B and InternVL-1.5 in Tab. 3, 4, and 5. The evaluation of InternLM2-20B is conducted based on their official codebase. We omit the images within the context and only evaluate InternLM2-20B on tasks with text needles. Note that InternLM2-20B is the language model used to initialize InternVL-1.5. According to the results in the tables, we can observe that InternLM2-20B and InternVL-1.5 achieve comparable performance when the context length is short. However, when the context length is larger than 32K, the performance of InternVL-1.5 is much inferior to InternLM2-20B, demonstrating that using only samples with a maximum context length of less than 4096 for multimodal fine-tuning can impair the model’s ability to handle long contexts. It is worth noting that InternLM2-20B also performs poorly in counting and reasoning tasks, which are more complex than retrieval. This phenomenon is consistent with the conclusions presented in RULER [84], which argues that despite achieving nearly perfect performance on the vanilla NIAH test, almost all models exhibit large degradation on more complex tasks as sequence length increases.

A.2 Qualitative Analyses

In this section, we present some error cases of InternVL-1.5 [17] in Fig. 5. As discussed in Appendix A.1, the examples in the first row show that placing questions before context fails to improve

Table 3: **Comparison of InternLM2 and InternVL-1.5 in text retrieval task.**

Model	**1∼2K**	**2∼4K**	**4∼8K**	**8∼12K**	**12∼16K**	**16∼24K**	**24∼32K**	**32∼40K**	**40∼48K**	**48∼64K**	**64∼72K**
InternLM2	98.4	99.4	99.2	97.3	95.3	93.1	91.9	90.7	88.2	82.4	82.7
InternVL-1.5	99.0	99.7	96.3	95.1	92.3	90.9	90.6	81.0	81.3	79.7	72.7

Table 4: **Comparison of InternLM2 and InternVL-1.5 in text counting task.**

Model	**1∼2K**	**2∼4K**	**4∼8K**	**8∼12K**	**12∼16K**	**16∼24K**	**24∼32K**	**32∼40K**	**40∼48K**	**48∼64K**	**64∼72K**
InternLM2	79.5	68.0	58.2	46.7	38.6	32.0	22.4	12.0	9.8	8.8	5.0
InternVL-1.5	67.6	60.0	46.7	46.8	33.3	28.0	17.0	8.3	5.4	7.7	6.8

Table 5: **Comparison of InternLM2 and InternVL-1.5 in text reasoning task.**

Model	**1∼2K**	**2∼4K**	**4∼8K**	**8∼12K**	**12∼16K**	**16∼24K**	**24∼32K**	**32∼40K**	**40∼48K**	**48∼64K**	**64∼72K**
InternLM2	88.1	83.0	81.0	68.7	69.0	60.9	50.6	39.4	40.1	34.0	31.9
InternVL-1.5	85.6	78.3	75.7	59.3	60.6	52.1	44.9	32.4	33.3	29.9	22.3

model performance but leading to the model collapse of InternVL-1.5 (*i.e.*, generating responses unrelated to the given question). Additionally, the examples in the second row demonstrate that InternVL-1.5 easily steps into a state of model collapse in the counting task, which means that the model tends to output a list that meets format requirements but lacks meaningful content, rather than accurately counting the number of needles in the text or images within the context. This phenomenon becomes more common when the given multimodal documents are exceptionally lengthy, causing the model to produce nonsensical text instead of answering the questions. Furthermore, the third row of Fig. 5 shows some error cases in the retrieval and reasoning tasks, where InternVL-1.5 struggles to understand the details of the images within the given context to answer these questions correctly. Based on these phenomena, we believe that increasing the model's instruction-following ability, improving its capacity to handle multiple images (so that it can at least accurately understand the number of images contained within the document), and enhancing its perception of image details are necessary steps to improve the model's ability for long multimodal document comprehension.

B Ethical discussion

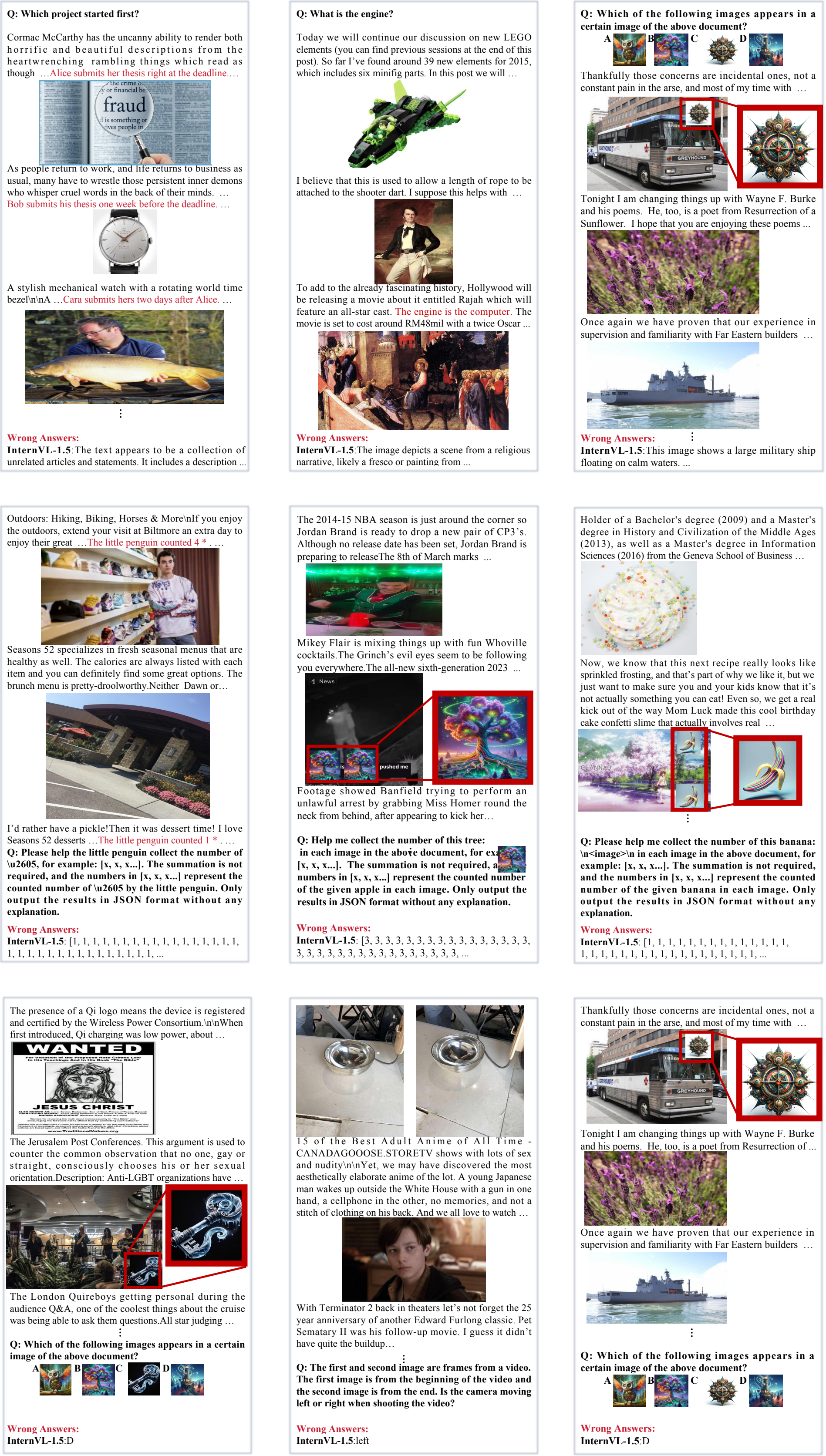
Our benchmark, Needle In A Multimodal Haystack (MM-NIAH), builds upon the OBELICS dataset, which has undergone extensive ethical review and content filtering to ensure compliance with ethical standards. The creation of OBELICS was guided by ethical principles, including respect for content creators' consent decisions and significant efforts to filter inappropriate content, such as pornographic material. Based on this solid foundation, all new contents (*i.e.*, text and image needles) introduced in MM-NIAH are carefully designed and manually verified, ensuring that the benchmark aligns with ethical guidelines and avoids the inclusion of any unreasonable or harmful content.

C License and Author Statement

We release the benchmark under the CC-BY license and Terms of Use. It is required to disclose the utilization when this benchmark is used for model evaluation purposes. This license does not replace the licenses of the source materials, and any use of content included in the dataset must comply with the original licenses and applicable rights of its data subjects. The purpose of this statement is to clarify the responsibilities and liabilities associated with the utilization of this benchmark. While we have spared no effort to ensure accuracy and legality of samples in our benchmark, we cannot guarantee its absolute completeness or correctness. Therefore, if any rights, legal or otherwise, are violated through this benchmark, including but not limited to copyright infringement, privacy violations, or misuse of sensitive information, we, the authors, assume no liability for such violations.

By accessing, downloading, or using this benchmark, you agree to assume sole responsibility for any legal or other consequences resulting from its utilization. You also acknowledge your commitment to adhere to all relevant laws, regulations, and ethical guidelines governing its usage. Your acceptance of this statement and adherence to the terms and conditions of the CC-BY license are implicit in your access, download, or utilization of this benchmark. If you do not agree with the terms outlined herein or the CC-BY license, you are not authorized to use this benchmark.

The benchmark will be hosted and maintained on Github and the Hugging Face Hub platform.



fraud
GREYHOUND
WANTED
For Violation of the Proposed Hate Crimes Law In His Teachings And In His Book "The Bible"
JESUS CHRIST
www.TraditionalValues.org
News
is
pushed me
CLANNAD
A
B
C
D

Q: Which project started first?

Cormac McCarthy has the uncanny ability to render both horrific and beautiful descriptions from the heartwrenching rambling things which read as though …Alice submits her thesis right at the deadline.…

As people return to work, and life returns to business as usual, many have to wrestle those persistent inner demons who whisper cruel words in the back of their minds. … Bob submits his thesis one week before the deadline. …

A stylish mechanical watch with a rotating world time bezel\n\nA …Cara submits hers two days after Alice. …

⋮

Wrong Answers:
InternVL-1.5:The text appears to be a collection of unrelated articles and statements. It includes a description ...

Q: What is the engine?

Today we will continue our discussion on new LEGO elements (you can find previous sessions at the end of this post). So far I've found around 39 new elements for 2015, which includes six minifig parts. In this post we will …

I believe that this is used to allow a length of rope to be attached to the shooter dart. I suppose this helps with …

To add to the already fascinating history, Hollywood will be releasing a movie about it entitled Rajah which will feature an all-star cast. The engine is the computer. The movie is set to cost around RM48mil with a twice Oscar ...

Wrong Answers:
InternVL-1.5:The image depicts a scene from a religious narrative, likely a fresco or painting from ...

Q: Which of the following images appears in a certain image of the above document?
A B C D

Thankfully those concerns are incidental ones, not a constant pain in the arse, and most of my time with …

Tonight I am changing things up with Wayne F. Burke and his poems. He, too, is a poet from Resurrection of a Sunflower. I hope that you are enjoying these poems ...

Once again we have proven that our experience in supervision and familiarity with Far Eastern builders …

⋮

Wrong Answers:
InternVL-1.5:This image shows a large military ship floating on calm waters. ...

Outdoors: Hiking, Biking, Horses & More\nIf you enjoy the outdoors, extend your visit at Biltmore an extra day to enjoy their great …The little penguin counted 4 * . …

Seasons 52 specializes in fresh seasonal menus that are healthy as well. The calories are always listed with each item and you can definitely find some great options. The brunch menu is pretty-droolworthy.Neither Dawn or…

I'd rather have a pickle!Then it was dessert time! I love Seasons 52 desserts …The little penguin counted 1 * . …

Q: Please help the little penguin collect the number of \u2605, for example: [x, x, x...]. The summation is not required, and the numbers in [x, x, x...] represent the counted number of \u2605 by the little penguin. Only output the results in JSON format without any explanation.

Wrong Answers:
InternVL-1.5: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...

The 2014-15 NBA season is just around the corner so Jordan Brand is ready to drop a new pair of CP3's. Although no release date has been set, Jordan Brand is preparing to releaseThe 8th of March marks ...

Mikey Flair is mixing things up with fun Whoville cocktails.The Grinch's evil eyes seem to be following you everywhere.The all-new sixth-generation 2023 ...

Footage showed Banfield trying to perform an unlawful arrest by grabbing Miss Homer round the neck from behind, after appearing to kick her…

Q: Help me collect the number of this tree:
in each image in the above document, for ex[x, x, x...]. The summation is not required, a numbers in [x, x, x...] represent the counted number of the given apple in each image. Only output the results in JSON format without any explanation.

Wrong Answers:
InternVL-1.5: [3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...

Holder of a Bachelor's degree (2009) and a Master's degree in History and Civilization of the Middle Ages (2013), as well as a Master's degree in Information Sciences (2016) from the Geneva School of Business …

Now, we know that this next recipe really looks like sprinkled frosting, and that's part of why we like it, but we just want to make sure you and your kids know that it's not actually something you can eat! Even so, we get a real kick out of the way Mom Luck made this cool birthday cake confetti slime that actually involves real …

⋮

Q: Please help me collect the number of this banana: \n<image>\n in each image in the above document, for example: [x, x, x...]. The summation is not required, and the numbers in [x, x, x...] represent the counted number of the given banana in each image. Only output the results in JSON format without any explanation.

Wrong Answers:
InternVL-1.5: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...

The presence of a Qi logo means the device is registered and certified by the Wireless Power Consortium.\n\nWhen first introduced, Qi charging was low power, about …

The Jerusalem Post Conferences. This argument is used to counter the common observation that no one, gay or straight, consciously chooses his or her sexual orientation.Description: Anti-LGBT organizations have …

The London Quireboys getting personal during the audience Q&A, one of the coolest things about the cruise was being able to ask them questions.All star judging …

⋮

Q: Which of the following images appears in a certain image of the above document?
A B C D

Wrong Answers:
InternVL-1.5:D

15 of the Best Adult Anime of All Time - CANADAGOOOSE.STORETV shows with lots of sex and nudity\n\nYet, we may have discovered the most aesthetically elaborate anime of the lot. A young Japanese man wakes up outside the White House with a gun in one hand, a cellphone in the other, no memories, and not a stitch of clothing on his back. And we all love to watch …

With Terminator 2 back in theaters let's not forget the 25 year anniversary of another Edward Furlong classic. Pet Sematary II was his follow-up movie. I guess it didn't have quite the buildup…

⋮

Q: The first and second image are frames from a video. The first image is from the beginning of the video and the second image is from the end. Is the camera moving left or right when shooting the video?

Wrong Answers:
InternVL-1.5:left

Thankfully those concerns are incidental ones, not a constant pain in the arse, and most of my time with …

Tonight I am changing things up with Wayne F. Burke and his poems. He, too, is a poet from Resurrection of ...

Once again we have proven that our experience in supervision and familiarity with Far Eastern builders …

⋮

Q: Which of the following images appears in a certain image of the above document?
A B C D

Wrong Answers:
InternVL-1.5:D

Figure 5: **Some error cases of InternVL-1.5.**

Table 6: **Results on Retrieval-Text-Needle.**

Model	1K	2K	4K	8K	12K	16K	24K	32K	40K	48K	64K	Overall
Emu2-Chat	65.3	54.3	18.6	3.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12.9
VILA-13B	93.7	86.6	59.2	38.5	15.2	6.8	0.9	0.0	0.7	0.0	0.0	27.4
IDEFICS2	95.0	90.7	31.8	11.8	15.1	3.0	0.0	0.0	0.0	0.0	0.0	22.5
LLaVA-1.6-13B	96.4	91.0	68.4	39.2	18.3	6.8	2.3	0.4	0.6	0.0	0.0	29.4
LLaVA-1.6-34B	98.5	96.5	89.9	77.3	53.8	4.3	0.0	0.0	0.0	0.0	0.0	38.2
InternVL-1.5	99.0	99.7	96.3	95.1	92.3	90.9	90.6	81.0	81.3	79.7	72.7	89.0
InternVL-1.5-RAG	99.4	99.6	99.1	99.0	98.0	96.5	96.3	96.1	94.1	95.3	94.9	97.1
Gemini-1.5	92.8	89.6	89.2	89.5	87.3	85.0	87.9	86.8	87.1	86.0	90.7	88.4
GPT-4V	97.5	98.2	95.6	96.0	100.0	100.0	95.6	96.0	76.0	92.5	95.0	94.8
Human	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 7: **Results on Counting-Text-Needle.**

Model	1K	2K	4K	8K	12K	16K	24K	32K	40K	48K	64K	Overall
Emu2-Chat	3.2	0.8	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4
VILA-13B	25.5	15.4	14.8	11.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0	6.1
IDEFICS2	42.6	15.6	1.8	1.2	1.4	0.0	0.0	0.0	0.0	0.0	0.0	5.7
LLaVA-1.6-13B	33.7	32.4	30.6	33.6	21.1	6.5	1.6	0.2	0.3	0.0	0.0	14.6
LLaVA-1.6-34B	55.0	47.6	34.8	19.2	3.2	0.0	0.0	0.0	0.0	0.0	0.0	14.5
InternVL-1.5	67.6	60.0	46.7	46.8	33.3	28.0	17.0	8.3	5.4	7.7	6.8	29.8
InternVL-1.5-RAG	80.7	70.4	52.3	52.9	57.8	52.7	40.7	36.6	28.5	19.5	12.4	45.9
Gemini-1.5	90.4	85.9	82.5	79.0	79.5	79.1	75.4	71.2	70.1	74.1	77.0	78.6
GPT-4V	70.0	90.4	84.7	84.1	82.2	72.8	73.6	64.6	55.6	53.6	77.6	73.6
Human	100.0	98.7	98.7	100.0	98.7	100.0	100.0	100.0	99.0	100.0	97.9	99.4

Table 8: **Results on Reasoning-Text-Needle.**

Model	1K	2K	4K	8K	12K	16K	24K	32K	40K	48K	64K	Overall
Emu2-Chat	48.7	47.5	31.1	12.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12.7
VILA-13B	64.9	51.9	47.4	35.6	24.5	5.2	1.2	0.0	0.0	0.0	0.7	21.0
IDEFICS2	73.6	48.1	17.1	11.7	10.1	1.2	0.0	0.0	0.0	0.0	0.0	14.7
LLaVA-1.6-13B	57.4	42.6	46.7	33.2	19.4	11.3	2.0	1.5	0.0	0.0	0.0	19.5
LLaVA-1.6-34B	76.5	69.7	61.8	43.6	27.8	4.6	0.0	0.0	0.0	0.0	0.0	25.8
InternVL-1.5	85.6	78.3	75.7	59.3	60.6	52.1	44.9	32.4	33.3	29.9	22.3	52.2
InternVL-1.5-RAG	89.4	86.6	79.2	66.4	63.8	69.4	63.9	61.0	64.1	59.0	58.9	69.3
Gemini-1.5	95.0	87.9	84.6	87.6	83.1	74.4	78.6	72.5	70.3	66.5	70.9	79.2
GPT-4V	95.6	93.5	89.8	93.3	79.8	79.3	65.0	98.0	76.0	76.1	76.7	83.9
Human	100.0	98.0	98.4	97.7	100.0	98.4	100.0	100.0	100.0	97.5	97.7	98.9

Table 9: **Results on Retrieval-Image-Needle.**

Model	1K	2K	4K	8K	12K	16K	24K	32K	40K	48K	64K	Overall
Emu2-Chat	26.3	23.6	14.8	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.9
VILA-13B	28.8	29.1	31.1	24.7	29.8	9.6	0.0	0.0	0.0	0.0	0.0	13.9
IDEFICS2	26.7	21.5	22.0	22.6	23.8	0.3	0.0	0.0	0.0	0.0	0.0	10.6
LLaVA-1.6-13B	32.2	34.6	26.6	26.7	24.1	23.9	6.0	0.0	0.0	0.0	0.0	15.8
LLaVA-1.6-34B	57.3	51.5	43.4	34.6	23.1	9.8	0.0	0.0	0.0	0.0	0.0	20.0
InternVL-1.5	25.0	24.4	26.4	26.2	33.1	31.4	31.4	28.5	25.2	30.6	26.4	28.0
InternVL-1.5-RAG	24.7	30.1	32.6	36.4	27.2	27.3	24.2	31.8	20.0	15.8	16.0	26.0
Gemini-1.5	17.9	17.7	22.7	23.5	25.9	26.4	27.7	20.8	21.6	19.6	22.2	22.4
Human	100.0	97.8	98.0	96.4	97.8	97.8	100.0	97.8	100.0	95.8	97.3	98.1

Table 10: **Results on Counting-Image-Needle.**

Model	1K	2K	4K	8K	12K	16K	24K	32K	40K	48K	64K	Overall
Emu2-Chat	0.0	0.0	1.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
VILA-13B	0.0	3.9	5.6	6.7	7.1	1.9	0.0	0.0	0.0	0.0	0.0	2.3
IDEFICS2	0.0	0.0	0.0	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
LLaVA-1.6-13B	12.0	20.2	31.7	23.1	12.3	5.5	1.0	0.0	0.2	0.4	0.0	9.7
LLaVA-1.6-34B	1.3	0.3	0.4	1.1	6.0	1.2	0.0	0.0	0.0	0.0	0.0	0.9
InternVL-1.5	30.6	16.6	6.1	0.7	0.5	0.3	0.0	0.0	0.0	0.0	0.0	5.0
InternVL-1.5-RAG	44.8	21.8	4.9	1.8	0.6	0.2	0.0	0.0	0.0	0.0	0.0	6.7
Gemini-1.5	52.1	29.8	17.0	10.4	6.9	8.3	6.0	6.3	5.0	3.6	6.4	13.8
Human	98.2	100.0	94.2	100.0	98.6	96.4	99.2	98.8	98.6	98.0	98.1	98.2

Table 11: **Results on Reasoning-Image-Needle.**

Model	1K	2K	4K	8K	12K	16K	24K	32K	40K	48K	64K	Overall
Emu2-Chat	54.3	40.9	37.2	17.6	5.1	0.0	0.0	0.0	0.0	0.0	0.0	14.1
VILA-13B	55.6	49.0	51.4	53.1	55.1	30.0	4.8	1.1	0.0	0.0	0.0	27.3
IDEFICS2	49.8	27.1	25.6	35.3	35.3	2.7	0.0	0.0	0.0	0.0	0.0	16.0
LLaVA-1.6-13B	50.1	49.2	45.7	54.1	50.7	39.1	21.0	2.4	0.0	0.0	0.0	28.4
LLaVA-1.6-34B	58.8	55.4	52.2	55.7	48.1	29.2	0.0	0.0	0.0	0.0	0.0	27.2
InternVL-1.5	49.2	52.8	49.5	50.1	51.3	48.5	53.2	48.9	44.4	51.1	52.2	50.1
InternVL-1.5-RAG	65.9	58.3	51.5	50.8	56.4	62.7	52.1	51.4	55.9	51.2	52.0	55.3
Gemini-1.5	39.6	38.9	45.1	52.3	49.7	49.1	45.7	53.7	60.9	54.1	54.3	49.4
Human	100.0	100.0	98.0	100.0	95.7	100.0	100.0	100.0	97.5	100.0	100.0	99.2

D Datasheet for MM-NIAH benchmark

D.1 Motivation

Q1 **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

- MM-NIAH was created to systematically evaluate the capability of existing Multimodal Large Language Models (MLLMs) to comprehend long multimodal documents, which is crucial for real-world applications but remains underexplored.

Q2 **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

- This dataset is presented by OpenGVLab of Shanghai AI Laboratory.

Q3 **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

- This work was sponsored by Shanghai AI Laboratory.

Q4 **Any other comments?**

- No.

D.2 Composition

Q5 **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

- Each instance in MM-NIAH represents a long multimodal document composed of interleaved image-text sequences and a corresponding question-anser pair.

Q6 **How many instances are there in total (of each type, if appropriate)?**

- The MM-NIAH benchmark comprises about 12,000 samples in total. Each type of evaluation data (retrieval, counting, reasoning) contains approximately 2,800 samples, with an equal distribution between text needles and image needles.

Q7 **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

- The dataset is created based on the interleaved image-text sequences from the OBELICS dataset. It includes a wide range of long multimodal documents to cover diverse scenarios for the evaluation tasks.

Q8 **What data does each instance consist of?** *"Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.*

- Each instance consists of a long multimodal document along with inserted needles (either text or image) for evaluation purposes.

Q9 **Is there a label or target associated with each instance?** *If so, please provide a description.*

- Yes, each instance has associated questions related to the inserted needles, which serve as the targets for the evaluation tasks.

Q10 **Is any information missing from individual instances?** *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

- No.

Q11 **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** *If so, please describe how these relationships are made explicit.*

- No.

Q12 **Are there recommended data splits (e.g., training, development/validation, testing)?** *If so, please provide a description of these splits, explaining the rationale behind them.*

- Yes, we provide validation split and test split. Note that the ground truth of the samples in the test split is not publicly available.

Q13 **Are there any errors, sources of noise, or redundancies in the dataset?** *If so, please provide a description.*

- We have conducted a quality check on this benchmark. However, due to the large volume of data, there may be a very small number of errors or omissions.

Q14 **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** *If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

- The benchmark is built upon the OBELICS dataset for the interleaved image-text sequences. There are no additional external resources required. The MM-NIAH benchmark includes all necessary data for evaluation purposes.

Q15 **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** *If so, please provide a description.*

- MM-NIAH is built upon the OBELICS dataset, which has undergone extensive ethical review and content filtering to ensure compliance with ethical standards. The creation of OBELICS was guided by ethical principles, including respect for content creators' consent decisions and significant efforts to filter inappropriate content, such as pornographic material. Based on this solid foundation, all new contents (*i.e.*, text and image needles) introduced in MM-NIAH are carefully designed and manually verified, ensuring that the benchmark aligns with ethical guidelines and avoids the inclusion of any unreasonable or harmful content.

Q16 **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** *If so, please describe why.*

- See Q15.

Q17 **Does the dataset relate to people?** *If not, you may skip the remaining questions in this section.*

- People might be found in the images or textual descriptions, but they are not the primary emphasis of the dataset.

Q18 **Does the dataset identify any subpopulations (e.g., by age, gender)?**

- We don't include any indicators of subpopulations as attributes.

Q19 **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** *If so, please describe how.*

- Yes, it may be possible to identify people using face recognition. We do not provide any such means nor make attempts, but institutions owning large amounts of face identifiers may identify specific people in the dataset. Similarly, people may be identified through the associated text.

Q20 **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** *If so, please provide a description.*

- See Q15.

Q21 **Any other comments?**

- No.

D.3 Collection Process

Q22 **How was the data associated with each instance acquired?** *Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

- We concatenate multiple interleaved image-text documents from OBELICS into a long-context document containing 1k to 72k image and text tokens. After that, we inject needles containing key information into a certain depth of the text or certain images within the document. To cover both text and image modalities, the proposed MM-NIAH comprises two types of needles (*i.e.*, text needles and image needles), where the needles inserted into the text are termed text needles while those inserted into images are termed image needles. All text and image needles used in MM-NIAH are manually verified to ensure the correctness.

Q23 **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** *How were these mechanisms or procedures validated?*

- We only use CPUs to generate these evaluation data. We validate our implementation by manually verifying all needles to be inserted and checking a subset of the generated evaluation data.

Q24 **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

- MM-NIAH is built upon the OBELICS dataset. We sample a subset of interleaved image-text sequences from it to construct the multimodal documents for evaluation purposes.

Q25 **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

- No crowdworkers were used in the curation of the dataset. Authors of this paper enabled its creation for no payment.

Q26 **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?** *If not, please describe the timeframe in which the data associated with the instances was created.*

- The licensed photos vary in their date taken over a wide range of years up to 2023.

Q27 **Were any ethical review processes conducted (e.g., by an institutional review board)?** *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

- As described in Q15, our benchmark is built up on the OBELICS dataset, which has undergone extensive ethical review and content filtering to ensure compliance with ethical standards. Based on this solid foundation, all new contents (*i.e.*, text and image needles) introduced in MM-NIAH are carefully designed and manually verified, ensuring that the benchmark aligns with ethical guidelines and avoids the inclusion of any unreasonable or harmful content. Therefore, we did not conduct a formal ethical review process via institutional review boards.

Q28 **Does the dataset relate to people?** *If not, you may skip the remaining questions in this section.*

- People might be present in the images and descriptions, although they are not the sole emphasis of the dataset.

Q29 **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

- To collect the data, we concatenate multiple interleaved image-text documents from OBELICS into a long-context document containing 1k to 72k image and text tokens. After that, we inject needles containing key information into a certain depth of the text or certain images within the document. To cover both text and image modalities, the proposed MM-NIAH comprises two types of needles (*i.e.*, text needles and image needles), where the needles inserted into the text are termed text needles while those inserted into images are termed image needles. All needles inserted into documents are manually verified.

Q30 **Were the individuals in question notified about the data collection?** *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

- Individuals were not notified about the data collection. Our benchmark is built upon the OBELICS dataset, which only contains information that is publicly available on the Internet. The publishers of these information are usually aware that it will be made public to the world, but they may not be aware that it will be collected in this way.

Q31 **Did the individuals in question consent to the collection and use of their data?** *If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

- No. See Q30.

Q32 **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

- Users can contact us to remove any annotation in our proposed MM-NIAH.

Q33 **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

- No. See Q27.

Q34 **Any other comments?**

- No.

D.4 Preprocessing, Cleaning, and/or Labeling

Q35 **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** *If so, please provide a description. If not, you may skip the remainder of the questions in this section.*

- MM-NIAH is established using the method described in Q22. All needles to be inserted into the multimodal documents are manually verified to ensure the correctness. Besides, a subset of evaluation data are sampled to be checked to further ensure the correctness.

Q36 **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** *If so, please provide a link or other access point to the "raw" data.*

- No.

Q37 **Is the software used to preprocess/clean/label the instances available?** *If so, please provide a link or other access point.*

- No.

Q38 **Any other comments?**

- No.

D.5 Uses

Q39 **Has the dataset been used for any tasks already?** *If so, please provide a description.*

- Only this paper used this benchmark for experiments up to date.

Q40 **Is there a repository that links to any or all papers or systems that use the dataset?** *If so, please provide a link or other access point.*

- Yes, we will maintain the leaderboard on the project page.

Q41 **What (other) tasks could the dataset be used for?**

- The dataset could be used to evaluate the comprehension ability for long multimodal documents.

Q42 **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** *For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

- No.

Q43 **Are there tasks for which the dataset should not be used?** *If so, please provide a description.*

- Our dataset should only be used for non-commercial academic research.

Q44 **Any other comments?**

- No.

D.6 Distribution

Q45 **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** *If so, please provide a description.*

- Yes, the benchmark will be open-sourced.

Q46 **How will the dataset be distributed (e.g., tarball on website, API, GitHub)?** *Does the dataset have a digital object identifier (DOI)?*

- The data will be available through GitHub.

Q47 **When will the dataset be distributed?**

- This benchmark is released at `https://github.com/OpenGVLab/MM-NIAH`.

Q48 **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

- CC-BY-4.0.

Q49 **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

- MM-NIAH owns the metadata and release as CC-BY-4.0.
- We do not own the copyright of the images.

Q50 **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

- No.

Q51 **Any other comments?**

- No.

D.7 Maintenance

Q52 **Who will be supporting/hosting/maintaining the dataset?**

- Huggingface will support hosting of the metadata.
- OpenGVLab of Shanghai AI Laboratory will maintain the samples distributed.

Q53 **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

- https://github.com/OpenGVLab/MM-NIAH

Q54 **Is there an erratum?** *If so, please provide a link or other access point.*

- Not at the moment. We plan to maintain it through GitHub issues and the README file.

Q55 **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

- No. However, specific samples can be removed on request.

Q56 **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** *If so, please describe these limits and explain how they will be enforced.*

- People may contact us to add specific samples to a blacklist.

Q57 **Will older versions of the dataset continue to be supported/hosted/maintained?** *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

- We will only support and maintain the latest version at all times and a new version release of MM-NIAH will automatically deprecate its previous version.

Q58 **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** *If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.*

- We welcome any contributions to MM-NIAH and we will announce updates regarding dataset extensions on GitHub. However, contributors must demonstrate the quality and harmlessness of the extended data annotations; otherwise, we will not accept these extensions.

Q59 **Any other comments?**

- No.