What Makes Multimodal In-Context Learning Work?

Supplementary Material

8. Appendix

8.1. Consideration on different behaviour of IDEFICS and OpenFlamingo

The two open-source models, IDEFICS [18] and OpenFlamingo [5], are both implementations of the model proposed by [3]. Despite sharing the same architecture, our analysis, as observable in 4 and 7 , reveals distinct behaviors between the two models when subjected to image removal or random image swapping. OpenFlamingo demonstrates a slight decrease in performance when removing or swapping images compared to the godlen prompt, indicating minimal impact from perturbations and recognising task, but not focusing on the image-text mapping. On the other hand, IDEFICS exhibits a larger performance drop without images and with random images experiences even further degradation with an increase in the number of shots.

dataset	num shots Prompt	4	8	16	32
Flickr30k	W/o image	61.11	63.45	62.57	61.66
	Rnd. image	51.04	53.15	58.20	59.07
	Base	60.92	62.42	64.05	63.03
ImageNet 1k	W/o image	25.67	24.05	20.93	16.43
	Rnd. image	11.09	7.73	6.16	5.18
	Base	22.55	21.54	18.73	16.11
MS-COCO	W/o image	83.33	88.68	93.36	94.50
	Rnd. image	76.31	84.89	90.55	NaN
	Base	84.43	91.34	96.52	NaN
rendered SST-2	W/o image	10.70	29.87	11.19	14.22
	Rnd. image	53.48	61.29	60.63	56.79
	Base	53.44	59.94	60.53	57.91

Table 4. Evaluation results using OpenFlamingo 9B and demonstrations sampled uniformly at random across four image-to-text datasets using 0, 4, 8, 16, and 32 in-context demonstrations. Depicted various prompt modifications, such as removing one modality (either the image or the question) or replacing it with a different random instance from the training dataset.

The disparity in behavior between the two models can likely be attributed to differences in their training datasets. IDEFICS was trained on the OBELICS [18] dataset, which contains longer, more contextual texts and extracts data directly from the HTML DOM tree, thus providing cleaner data free from ads and spam. This method ensures higher document quality, comparable to renowned datasets like The Pile and Wikipedia. Furthermore, OBELICS addresses the issue of image duplication present in Multimodal C4, in which only 60% of images are unique, thus offering a higher quality and more efficient training dataset. In contrast, OpenFlamingo was trained on the shorter, less detailed texts of Multimodal C4.

Given that IDEFICS generally achieves better scores and is more responsive to ICL, we have chosen to focus our study on this model.

Comparaison with Chen et al. [7] The findings presented by Chen et al. [7], corroborate the behavioral differences between the two models that we observed. However, their study emphasizes the behavior of OpenFlamingo and concludes that ICLis primarily driven by text, as it appears insensitive to changes in images. Our observations regarding VQA align with this: ICL indeed seems to be driven predominantly by text. However, we note a different pattern in image-to-text tasks, where ICL does respond to visual elements. Nonetheless, when text is also available, it tends to become the dominant factor influencing the model's responses.

8.2. Balanced sampling

In Section 5.1, we demonstrated that the performance of RICES ICL improves significantly due to a majority voting process that selects the most common label in a given context. To better understand how label imbalance impacts this, we conducted experiments in a binary classification framework, adjusting the sampling method to ensure an equal number of demonstrations from each class in the context. For random sampling, the demonstrations were arranged without specific order, while for RICES, we selected the closest demonstrations from each class and sorted them by increasing similarity. In Tab. 5, we found the following order of performance from worst to best: random sampling (comparaison point), balanced random sampling (+1.74% improvement), balanced RICES sampling (+8.40% improvement), and RICES sampling alone (+18.90% improvement). This suggests that while balancing the samples improves performance in random contexts, the balanced RICES approach yields only half the performance boost compared to using RICES alone. Therefore, we can infer that while example similarity contributes to model performance, the distribution of labels plays an important role.

dataset	num shots sampling	4	8	16	32
Hateful Memes	RICES	60.50	62.30	63.40	62.60
	Balanced rnd.	53.30	53.37	55.03	55.17
	Balanced RICES	54.60	56.10	58.30	57.70
	Random	50.57	50.93	52.00	53.77
rendered SST-2	RICES	75.80	84.14	82.84	80.18
	Balanced rnd.	57.07	57.27	58.11	58.85
	Balanced RICES	61.46	70.74	77.30	80.34
	Random	56.41	56.81	57.62	58.67

Table 5. Evaluation results using IDEFICS 9B across two binary classification vision-language datasets using 0, 4, 8, 16, and 32 in-context demonstrations. Depicted various sampling methods, random sampling (Random), RICES and their balanced counterparts.

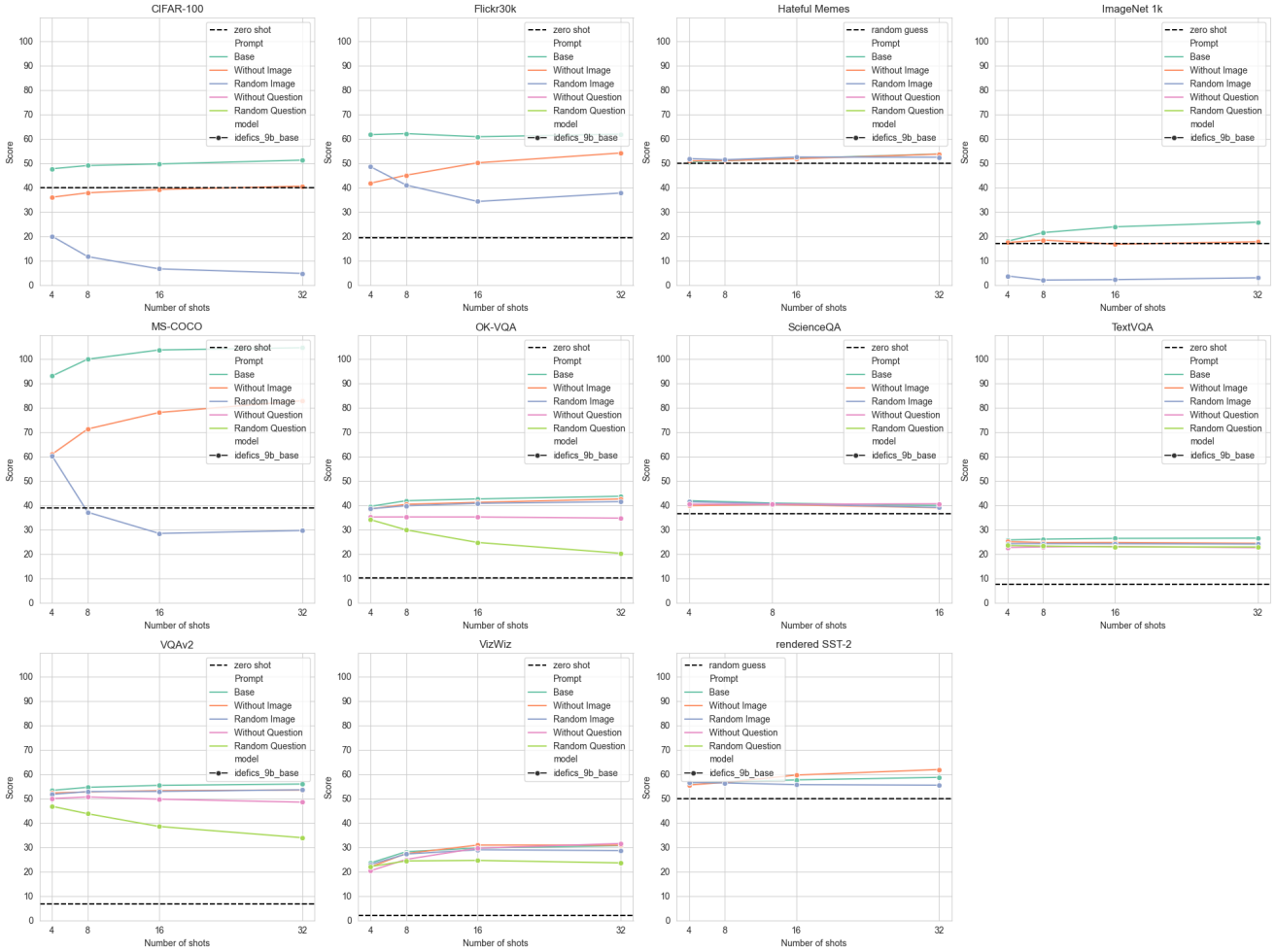

Figure 10. Full evaluation results using IDEFICS 9B and demonstrations sampled uniformly at random across twelve vision-language datasets using 0, 4, 8, 16, and 32 in-context demonstrations. Depicted various prompt modifications, such as removing one modality (either the image or the question) or replacing it with a different random instance from the training dataset.

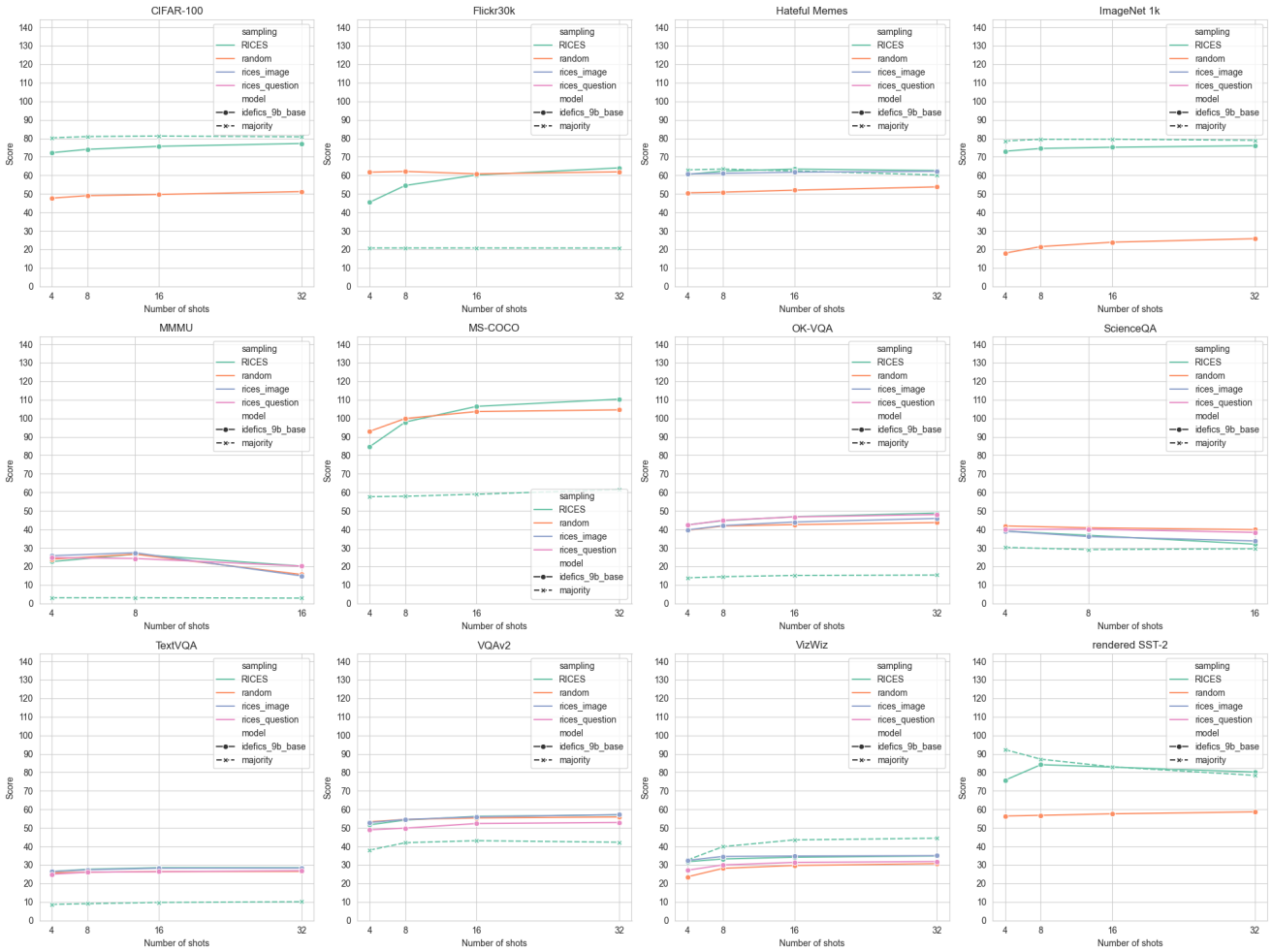

Figure 11. Full evaluation results using IDEFICS 9B and base prompt across twelve vision-language datasets using 0, 4, 8, 16, and 32 in-context demonstrations. Depicted the scores of random sampling (Random) and RICES in is standard form or using only one modality for similarity function (rices_modality)

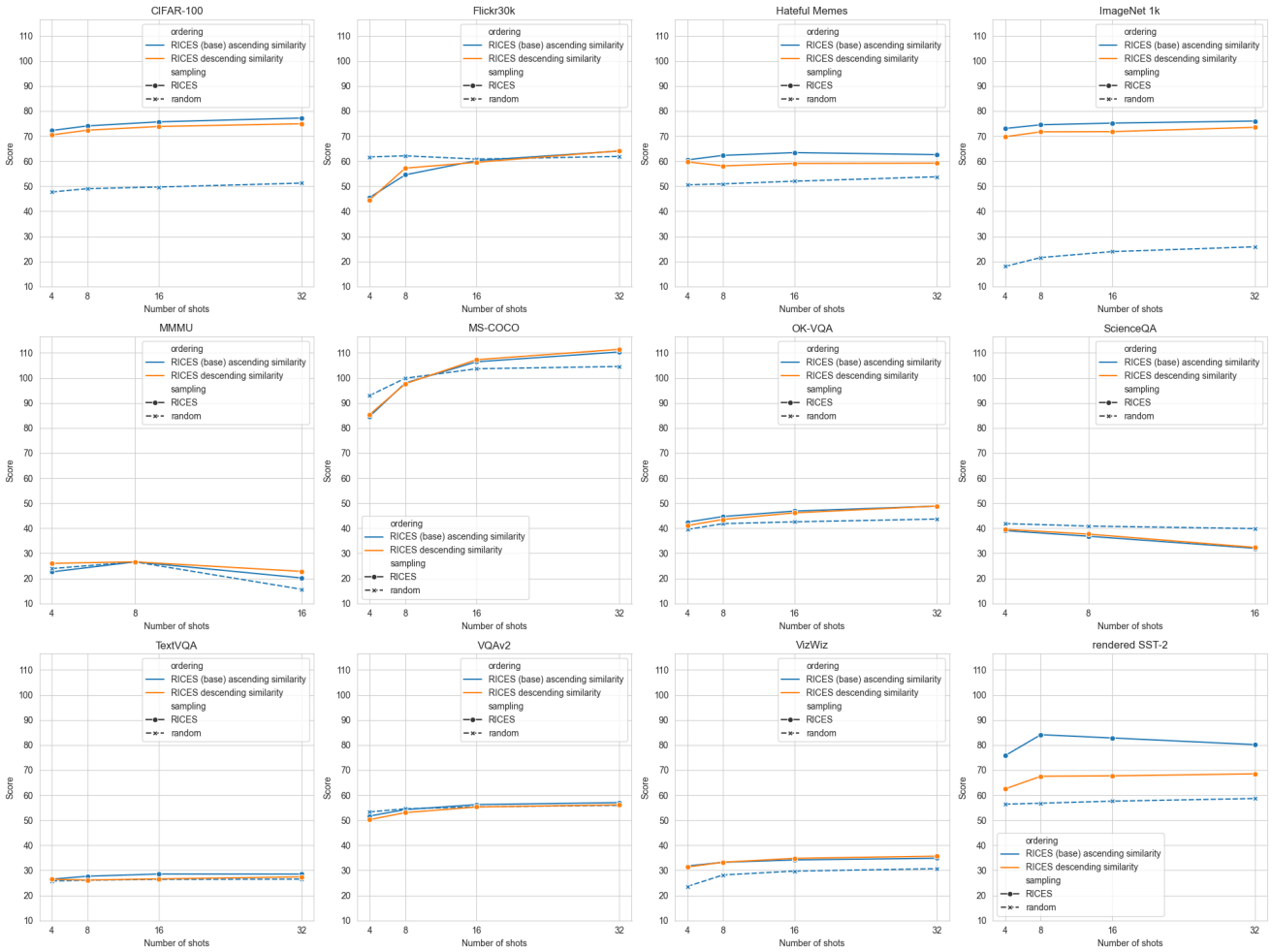

Figure 12. Evaluation results using IDEFICS 9B and base prompt across twelve vision-language datasets using 0, 4, 8, 16, and 32 in-context demonstrations. Comparison of RICES with default order of demonstration (ascending) and a variant with descending similarity ordering.

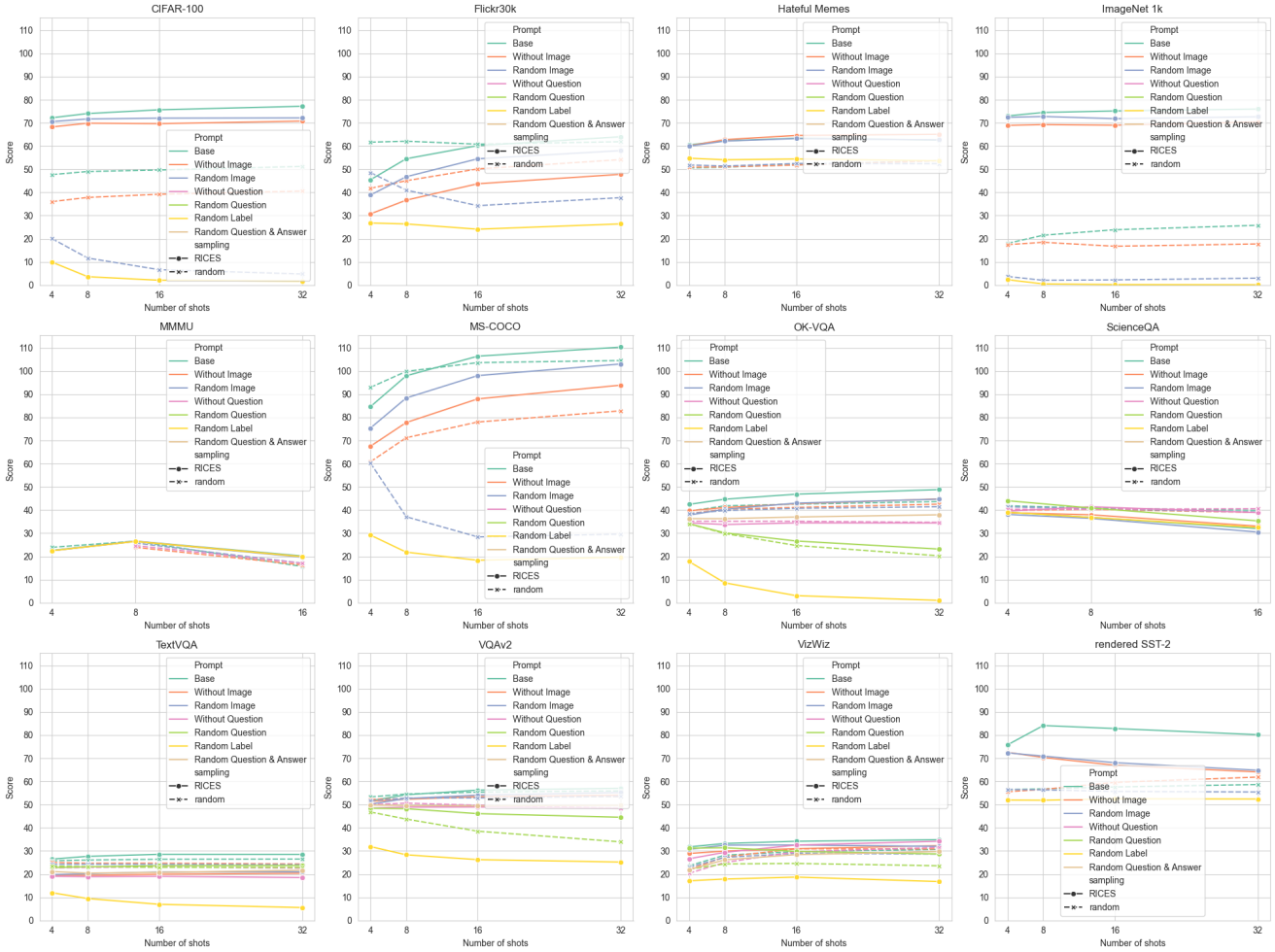

Figure 13. ull evaluation results using IDEFICS 9B and demonstrations sampled with RICES across twelve vision-language datasets using 0, 4, 8, 16, and 32 in-context demonstrations. Depicted various prompt modifications, such as removing one modality (either the image or the question) or replacing it with a different random instance from the training dataset.

Dataset	Shots Sampling	4	8	16	32
CIFAR-100	R. image	72.24	74.06	75.68	77.20
	Random	47.70	49.03	49.68	51.23
Flickr30k	RICES	45.48	54.54	60.21	64.03
	Random	61.69	62.12	60.83	61.89
Hateful Memes	RICES	60.50	62.30	63.40	62.60
	Random	50.57	50.93	52.00	53.77
	R. image	60.80	61.10	61.70	62.20
	R. OCR	59.80	61.70	62.30	62.80
ImageNet 1k	RICES	73.04	74.52	75.18	76.00
	Random	18.01	21.53	23.90	25.81
MMMU	RICES	22.60	26.60	20.20	NaN
	Random	23.93	26.60	15.67	NaN
	R. image	25.80	27.40	14.90	NaN
	R. question	24.80	24.30	20.20	NaN
MS-COCO	RICES	84.65	97.98	106.44	110.36
	Random	92.98	99.88	103.66	104.57
OK-VQA	RICES	42.47	44.70	46.87	48.84
	Random	39.54	41.85	42.58	43.68
	R. image	39.71	42.10	44.00	45.94
	R. question	42.35	44.92	46.74	48.02
ScienceQA	RICES	39.07	36.84	32.03	NaN
	Random	41.88	40.89	39.88	NaN
	R. image	39.07	36.14	33.76	NaN
	R. question	39.96	40.16	38.37	NaN
TextVQA	RICES	26.48	27.67	28.54	28.51
	Random	25.77	26.09	26.40	26.50
	R. image	26.39	27.38	28.24	28.33
	R. question	24.97	26.10	26.37	26.91
VQAv2	RICES	51.68	54.25	56.26	57.04
	Random	53.33	54.58	55.39	55.93
	R. image	52.92	54.57	55.98	57.20
	R. question	48.99	49.88	52.37	52.95
VizWiz	RICES	31.75	33.23	34.17	34.85
	Random	23.58	28.18	29.71	30.69
	R. image	32.45	34.61	34.82	35.03
	R. question	27.15	30.02	31.37	31.83
rendered SST-2	RICES	75.80	84.14	82.84	80.18
	Random	56.41	56.81	57.62	58.67

Table 6. Full evaluation results using IDEFICS 9B and base prompt across twelve vision-language datasets using 0, 4, 8, 16, and 32 in-context demonstrations. Depicted the scores of random sampling (Random) and RICES in is standard form or using only one modality for similarity function (R. modality)

Dataset	Shots Prompt	4	8	16	32
CIFAR-100	W/o image	36.03	37.84	39.21	40.57
	Rnd. image	20.05	11.67	6.65	4.79
	Base	47.70	49.03	49.68	51.23
Flickr30k	W/o image	41.78	45.05	50.15	54.16
	Rnd. image	48.54	40.99	34.29	37.75
	Base	61.69	62.12	60.83	61.89
Hateful Memes	W/o image	50.93	51.03	51.83	53.70
	Rnd. image	51.87	51.40	52.50	52.43
	Base	50.57	50.93	52.00	53.77
ImageNet 1k	W/o image	17.46	18.47	16.77	17.77
	Rnd. image	3.72	2.07	2.25	2.99
	Base	18.01	21.53	23.90	25.81
MS-COCO	W/o image	60.87	71.25	78.02	82.86
	Rnd. image	60.29	37.12	28.43	29.63
	Base	92.98	99.88	103.66	104.57
OK-VQA	W/o image	38.48	40.39	41.17	42.56
	W/o question	35.18	35.19	35.17	34.70
	Rnd. image	38.54	39.78	40.77	41.46
	Rnd. question	34.06	29.88	24.72	20.27
	Base	39.54	41.85	42.58	43.68
ScienceQA	W/o image	39.83	40.31	38.99	NaN
	W/o question	40.41	40.37	40.59	NaN
	Rnd. image	41.41	40.64	39.12	NaN
	Base	41.88	40.89	39.88	NaN
TextVQA	W/o image	25.08	24.69	24.71	24.38
	W/o question	22.66	22.90	23.08	22.58
	Rnd. image	24.25	24.26	24.22	24.08
	Rnd. question	23.49	23.23	22.92	22.87
	Base	25.77	26.09	26.40	26.50
VQAv2	W/o image	52.26	52.67	53.22	53.47
	W/o question	49.98	50.63	49.73	48.49
	Rnd. image	51.67	52.84	52.90	53.57
	Rnd. question	46.80	43.75	38.52	33.92
	Base	53.33	54.58	55.39	55.93
VizWiz	W/o image	21.96	27.44	30.90	30.89
	W/o question	20.36	25.02	29.63	31.55
	Rnd. image	22.94	27.22	28.97	28.65
	Rnd. question	22.08	24.40	24.60	23.59
	Base	23.58	28.18	29.71	30.69
rendered SST-2	W/o image	55.55	56.57	59.61	61.88
	Rnd. image	56.57	56.37	55.69	55.46
	Base	56.41	56.81	57.62	58.67

Table 7. Full evaluation results using IDEFICS 9B and demonstrations sampled uniformly at random across twelve vision-language datasets using 0, 4, 8, 16, and 32 in-context demonstrations. Depicted various prompt modifications, such as removing one modality (either the image or the question) or replacing it with a different random instance from the training dataset.

Dataset	Shots ordering	4	8	16	32
CIFAR-100	ascending	72.24	74.06	75.68	77.20
	descending	70.46	72.36	73.84	74.92
Flickr30k	ascending	45.48	54.54	60.21	64.03
	descending	44.56	57.21	59.53	64.11
Hateful Memes	ascending	60.50	62.30	63.40	62.60
	descending	59.70	58.10	59.10	59.20
ImageNet 1k	ascending	73.04	74.52	75.18	76.00
	descending	69.74	71.70	71.78	73.50
MMMU	ascending	22.60	26.60	20.20	NaN
	descending	26.00	26.60	22.80	NaN
MS-COCO	ascending	84.65	97.98	106.44	110.36
	descending	85.37	97.68	107.28	111.41
OK-VQA	ascending	42.47	44.70	46.87	48.84
	descending	41.09	43.54	46.16	48.88
ScienceQA	ascending	39.07	36.84	32.03	NaN
	descending	39.56	37.68	32.37	NaN
TextVQA	ascending	26.48	27.67	28.54	28.51
	descending	26.42	26.14	26.61	27.40
VQAv2	ascending	51.68	54.25	56.26	57.04
	descending	50.26	53.04	55.30	56.16
VizWiz	ascending	31.75	33.23	34.17	34.85
	descending	31.33	33.26	34.80	35.66
rendered SST-2	ascending	75.80	84.14	82.84	80.18
	descending	62.52	67.54	67.72	68.52

Table 8. Evaluation results using IDEFICS 9B and base prompt across twelve vision-language datasets using 0, 4, 8, 16, and 32 in-context demonstrations. Comparison of RICES with default order of demonstration (ascending) and a variant with descending similarity ordering.

Dataset	Shots Variant	4	8	16	32
CIFAR-100	Rnd. S. LMM	47.70	49.03	49.68	51.23
	RICES LMM	72.24	74.06	75.68	77.20
	RICES KNN	80.28	80.96	81.24	80.82
Flickr30k	Rnd. S. LMM	61.69	62.12	60.83	61.89
	RICES LMM	45.48	54.54	60.21	64.03
	RICES KNN	20.77	20.77	20.77	20.73
Hateful Memes	Rnd. S. LMM	50.57	50.93	52.00	53.77
	RICES LMM	60.50	62.30	63.40	62.60
	RICES KNN	63.00	63.40	62.40	60.20
ImageNet 1k	Rnd. S. LMM	18.01	21.53	23.90	25.81
	RICES LMM	73.04	74.52	75.18	76.00
	RICES KNN	78.58	79.46	79.52	78.90
MMMU	Rnd. S. LMM	23.93	26.60	15.67	NaN
	RICES LMM	22.60	26.60	20.20	NaN
	RICES KNN	3.10	3.10	2.90	NaN
MS-COCO	Rnd. S. LMM	92.98	99.88	103.66	104.57
	RICES LMM	84.65	97.98	106.44	110.36
	RICES KNN	57.69	57.90	59.00	61.55
OK-VQA	Rnd. S. LMM	39.54	41.85	42.58	43.68
	RICES LMM	42.47	44.70	46.87	48.84
	RICES KNN	13.86	14.46	15.14	15.35
ScienceQA	Rnd. S. LMM	41.88	40.89	39.88	NaN
	RICES LMM	39.07	36.84	32.03	NaN
	RICES KNN	30.29	29.10	29.55	NaN
TextVQA	Rnd. S. LMM	25.77	26.09	26.40	26.50
	RICES LMM	26.48	27.67	28.54	28.51
	RICES KNN	8.69	9.09	9.75	10.13
VQAv2	Rnd. S. LMM	53.33	54.58	55.39	55.93
	RICES LMM	51.68	54.25	56.26	57.04
	RICES KNN	38.01	42.01	43.12	42.25
VizWiz	Rnd. S. LMM	23.58	28.18	29.71	30.69
	RICES LMM	31.75	33.23	34.17	34.85
	RICES KNN	32.66	39.91	43.55	44.43
rendered SST-2	Rnd. S. LMM	56.41	56.81	57.62	58.67
	RICES LMM	75.80	84.14	82.84	80.18
	RICES KNN	92.26	87.12	82.96	78.38

Table 9. Evaluation results using IDEFICS 9B across twelve vision-language datasets using 0, 4, 8, 16, and 32 in-context demonstrations. Depicted M-ICL with random sampling (Rnd. S. LMM), M-ICL with RICES sampling (RICES LMM) and the majority voting baseline (RICES KNN)

Dataset	Shots / Prompt	4	8	16	32
CIFAR-100	W/o image	68.28	69.88	69.68	70.80
	Rnd. image	70.59	71.71	72.07	72.18
	Rnd. label	9.91	3.63	2.09	1.72
	Base	72.24	74.06	75.68	77.20
Flickr30k	W/o image	30.75	36.66	43.75	47.83
	Rnd. image	38.88	46.78	54.52	58.04
	Rnd. label	26.80	26.40	24.12	26.42
	Base	45.48	54.54	60.21	64.03
Hateful Memes	W/o image	60.10	62.80	64.60	65.10
	Rnd. image	60.00	62.17	63.27	62.70
	Rnd. label	54.77	54.10	54.43	53.67
	Base	60.50	62.30	63.40	62.60
ImageNet 1k	W/o image	68.94	69.28	69.02	70.42
	Rnd. image	72.41	72.79	71.84	72.67
	Rnd. label	2.32	0.51	0.23	0.14
	Base	73.04	74.52	75.18	76.00
MMMU	W/o image	22.40	26.40	19.90	NaN
	Rnd. image	22.47	26.47	19.67	NaN
	Rnd. label	22.47	26.47	19.90	NaN
	Base	22.60	26.60	20.20	NaN
MS-COCO	W/o image	67.58	77.81	88.01	93.93
	Rnd. image	75.42	88.40	98.06	103.08
	Rnd. label	29.19	21.85	18.34	19.64
	Base	84.65	97.98	106.44	110.36
OK-VQA	W/o image	39.69	41.11	42.76	44.82
	W/o quest.	34.50	33.77	34.47	34.44
	Rnd. image	37.83	40.37	43.01	44.72
	Rnd. label	17.80	8.60	3.06	1.02
	Rnd. quest.	34.11	30.20	26.66	23.18
	Base	42.47	44.70	46.87	48.84
ScienceQA	W/o image	38.7	37.98	33.07	NaN
	W/o quest.	39.56	41.45	38.92	NaN
	Rnd. image	38.13	36.42	30.52	NaN
	Rnd. label	39.07	36.84	32.52	NaN
	Rnd. quest.	44.04	40.92	35.35	NaN
	Base	39.07	36.84	32.03	NaN
TextVQA	W/o image	19.68	19.43	19.94	20.36
	W/o quest.	19.05	18.90	19.04	18.51
	Rnd. image	19.77	20.17	20.91	20.97
	Rnd. label	11.97	9.44	6.96	5.56
	Rnd. quest.	22.82	23.17	23.63	23.71
	Base	26.48	27.67	28.54	28.51
VQAv2	W/o image	51.43	52.32	53.39	54.07
	W/o quest.	48.47	48.90	48.96	48.38
	Rnd. image	50.21	52.63	54.08	55.22
	Rnd. label	31.87	28.32	26.24	25.20
	Rnd. quest.	48.42	48.24	46.10	44.54
	Base	51.68	54.25	56.26	57.04
VizWiz	W/o image	28.89	29.95	30.98	32.29
	W/o quest.	26.62	29.39	32.54	34.17
	Rnd. image	30.67	32.51	32.56	31.98
	Rnd. label	17.11	17.86	18.70	16.78
	Rnd. quest.	31.21	31.37	29.81	28.69
	Base	31.75	33.23	34.17	34.85
rendered SST-2	W/o image	72.56	70.36	67.00	64.10
	Rnd. image	72.35	70.93	68.11	64.83
	Rnd. label	51.97	51.89	52.53	52.41
	Base	75.80	84.14	82.84	80.18

Table 10. Full evaluation results using IDEFICS 9B and demonstrations sampled with RICES across twelve vision-language datasets using 0, 4, 8, 16, and 32 in-context demonstrations. Depicted various prompt modifications, such as removing one modality (either the image or the question) or replacing it with a different random instance from the training dataset.

Dataset	Zero-shot score
ScienceQA	36.39
MMMU	4.37
MS-COCO	38.94
Flickr30k	19.44
OK-VQA	10.29
VQAv2	6.66
VizWiz	2.16
ImageNet 1k	16.98
Hateful Memes	0.00
TextVQA	7.66
rendered SST-2	0.02
CIFAR-100	39.98

Table 11. Full evaluation results using IDEFICS 9B across twelve vision-language datasets using no demonstrations.

Dataset	Oracle RICES score
CIFAR-100	91.98
Flickr30k	76.53
Hateful Memes	100.00
ImageNet 1k	99.56
MMMU	19.30
MS-COCO	139.03
OK-VQA	75.90
ScienceQA	35.05
TextVQA	49.79
VQAv2	82.97
VizWiz	44.26
rendered SST-2	100.00

Table 12. Evaluation results using IDEFICS 9B and demonstrations sampled with RICES using ground truth as similarity across twelve vision-language datasets using 16 in-context demonstrations.