Supplementary material for "FETA: Towards Specializing Foundation Models for Expert Task Applications"

Amit Alfassy*[1,3] **Assaf Arbelle***[1] **Oshri Halimi**[1,3] **Sivan Harary**[1]
Roei Herzig[1] **Eli Schwartz**[1] **Rameswar Panda**[1] **Michele Dolfi**[1] **Christoph Auer**[1]
Kate Saenko[2,4] **Peter W. J. Staar**[1] **Rogerio Feris**[2] **Leonid Karlinsky***[2]

[1]IBM Research, [2]MIT-IBM AI-Watson Lab, [3]Technion, [4]Boston University

In these supplementary materials, we provide additional information about our data, baseline methods, implementation details, experimental results, and qualitative examples. Specifically,Section 6 provides additional implementation details about the data collection and processing pipelines, Section 3 provides additional details on the training procedure and runtime. In Section 4 we provide extensive an extensive ablation study, several additional results including results of the FLAVA [5] FM model and results of our CLIP-based models on the IKEA dataset. In Sections 5-7 we provide information about the code, data download and the license.

1 Data Collection Details

1.1 Data Download and Conversion

Car Manuals: The car manuals data was downloaded from https://www.workshopservicemanual.com. Table1 of the original paper details the amount of documents downloaded by car manufacturers. The documents contained an average of 149 pages per document.
IKEA Catalogs: The IKEA catalogs data was initially presented in [8] and downloaded from https://github.com/ivc-yz/SSR. The data in consisted of 29 IKEA US catalogs between the years 1986 and 2005, each document contain an average of 283 pages.

The downloaded documents were processed by the DeepSearch tool https://ds4sd.github.io/ which extracted the images and the texts. The extracted texts were further processed to create large, consecutive chunks of texts. We removed bad characters artifacts created during PDF parsing, we also filtered improbable boxes and failed boxes, finally we merged together spatially close text boxes.

The final result of our data conversion can be seen in Figures 5-7 for examples from the IKEA dataset and Figures 8-10 for examples from the Car-Manuals dataset. Red boxes mark extracted text blocks, blue boxes mark extracted images.

1.2 Identical Image Detection Process

We found that the car manual data includes images which reappear in several different locations within the same manual. Since our automatic annotation links images and texts which are located in the same page, retrieving a correct image but from a different page can artificially lower the test results. As our goal was to keep the flow unsupervised as much as possible, in order to overcome this issue we processed the images in three steps. First, we trained a self-supervised network on all the data in order to get meaningful image features. For that we used DINO [1], which we have found to

Submitted to the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks. Do not distribute.

create good image representations due to its loss function which inherently produces good clusters in the embedded space. As a second step, for each image we selected the top ten nearest neighbors in the embedded space. Lastly we performed Normalized Cross Correlation [6] filtering on the selected images and selected images with correlation higher than $t > 0.7$. These images were treated as identical images during test time. For the retrieval tests, the sets of texts matching to identical images were merged (by their union), and there was no penalty when retrieving an identical image from a different page. It is important to note that the DINO model used for identical images filtering was used only for that and not used in our experiments in any other way.

1.3 Text bounding box merging

OCR engines sometimes fail in parsing full paragraphs and end up splitting them to numerous bounding boxes. In order to lessen the effect such OCR errors has on FETA, we use a mechanism to merge adjacent bounding boxes into one bounding box. The process has 2 stages, in the first stage We employ a dilation technique in which we increase the length of each of the box's horizontal edges by a constant which is a percentage of the page's horizontal length (we used 1%), we also increase the length of each box's vertical edge by a bigger constant (4* times the horizontal constant) as text bounding boxes tend to be wide and short. This creates some overlaps between neighboring boxes. In the second stage we merge all the boxes which has any kind of overlap between them. Each group of overlapping boxes are merged into a bounding box that minimally contains the boxes being merged.

1.4 Manual Annotation Statistics

We created manual annotations for part of the Car Manuals dataset. Our manually labeled data consists of 15 documents and has 449 image-text pairs. We randomly selected these 15 documents, making sure to select at least two documents from each manufacturer. We then manually selected up to 50 images per document and manually generated (by a human expert) the image description using the information on the page as technical reference. This annotation was used to both validate the test results obtained on the automatically curated data and demonstrate that pre-training on automatically curated data indeed improves results on manually annotated data (Table 3 in the main paper). From Tables 2 and 3 in the main paper show that the same trends in terms of relative performance of different baselines appear both on the manual and the automatically curated data.

1.5 Further details on the automatic annotation

Figure 1 show the main steps in creating our automatic annotation of images - texts bags, the figure demonstrate the automatic process from a zip of PDFs to bags annotations which enable MIL training and test on the data. The steps are shown on an example pages from the cars dataset. The images above the flow chart show the creation of image-texts bags, while the images below the flow chart shot the creation of text-images bags.

2 Additional data analysis

2.1 Automatic vs manual text annotations cover

In order to asses the quality of the automatic annotations, we compare here between the manual annotations and automatic annotations. In this test we consider only the manually annotated documents. We report the percentage of the times, when looking at a specific image, the manually annotated text is contained within the automatically extracted texts for that same image (with a significant overlap). Over the cars data the cover is 93 percent. This high overlap is aligned with the Multiple Instance Learning setting where the MIL bag is assumed to contain at least one true sample. We consider the remainder or 7% as annotation noise. In the future versions of FETA it would be possible to increase this coverage by considering extracting text from adjacent pages, as well as using PDF sections parsing results.

2.2 Data quality check

We have asked three external reviewers to go over a subset of the automatic annotations. Each reviewer was asked to rate the annotation as good or bad, we present the findings of this experiment in Table 1.

Table 1: **Data quality check of automatic annotations**

	Good (out of 3)		
	1+	2+	3
Automatic annotations	93.6%	90.2%	82.3%

2.3 Data statistics and comparison to other datasets

In order to further demonstrate the difference and resemblance between cars and IKEA data sets to each and to other popular data sets, we show some data statistics in Table 2, all tokens data was calculated by us using CLIP's tokenizer. Word count to vocabulary size was calculated as the number of words occurrences in the entire dataset divided by a set of all unique words, present in the dataset, all number except CC3M(taken from original paper) were calculated by us. Image-page ratio reports the ratio between image area to entire page area in the dataset.

Table 2: **Data Statistics** In this table we compare the statistics of the texts in the Car-Manuals and IKEA datasets to common V&L datasets, i.e COCO, Flickr30K and CC3M

Measure	Cars	IKEA	Flickr30k	COCO	CC3M
Mean tokens per caption	44.2	59.1	15.7	13.5	10.3
Std tokens per caption	67.3	83.9	5.6	2.7	4.5
Unique tokens	11152	14942	15351	17624	51201
Average words per image	79.5	208.9	66.9	53	51.5
Word count to vocabulary size	48.3	14.3	87.4	97.2	804.8
Total images number	52119	9574	31783	82783	3318333

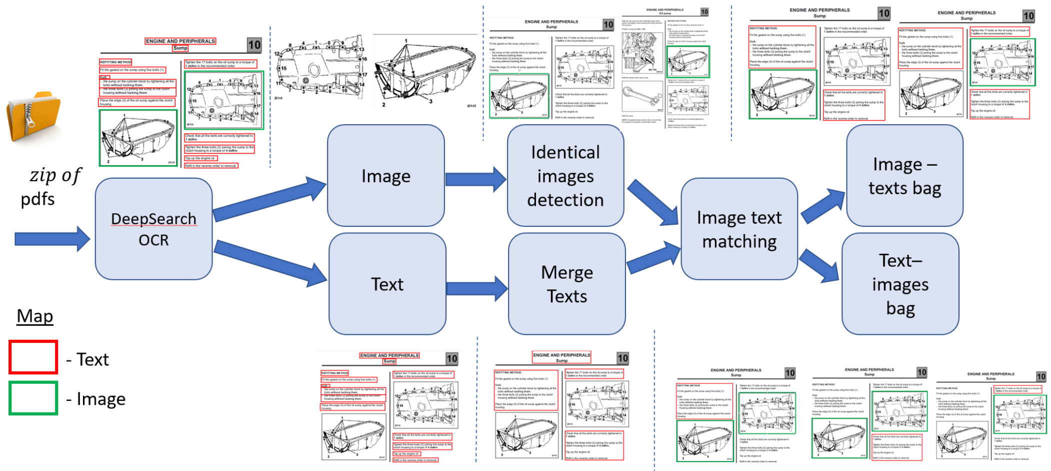


Figure 1: **Simplified figure explaining the automatic annotation process** showed using an example page from the cars dataset.

2.4 Common nouns and adjectives in data

To further understand the difference between the Car-Manuals and IKEA datasets to each other and to other popular datasets we report the most common nouns and adjectives present in each dataset sorted by appearance count. As can be seen in Figures 2 and 3 and in Tables 3 and 4, the nouns and adjectives from Flicker30K, COCO, CC3M are very similar and all three include nouns such as: person, beach etc. also, they all use the same adjectives such as white, young, etc. It seems that those 3 very popular datasets all reside in a very close domain and treat the same kind of popular data of natural images. When looking at the nouns and adjectives in Cars and IKEA, we see nouns such as: engine, connector, table, cotton which are specific to the expert domain each dataset deals with. We also see adjectives like, diagnostic, new, solid, adjustable which are again a strong characteristic of the expert domains of Cars and IKEA.The resemblance between Flickr30K, COCO and CC3M, coupled with the difference between them to our Cars and IKEA datasets, further strengthen our claim that FETA can indeed be useful in order to expand current FMs research to new under studied domains of expert V&L tasks which may provide noticeable value for practical real world applications.

Table 3: **Most common nouns by dataset** ordered by count from left to right.

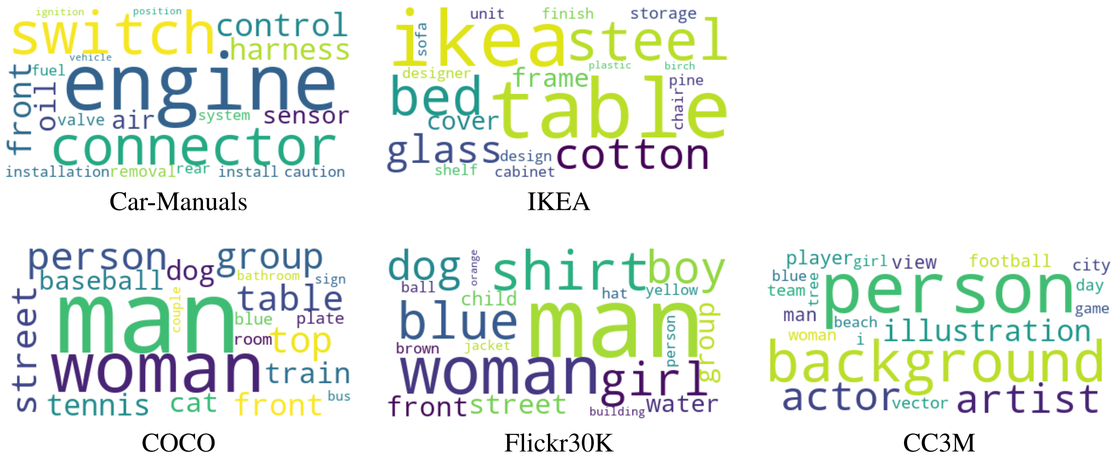
Dataset	Most Common Nouns
Cars	engine, switch, connector, control, front, harness, oil, air, fuel, installation, system, rear, caution, ignition, position, vehicle, side, ground, terminal, brake, door, unit, cylinder.
IKEA	table, steel, bed, cotton, glass, frame, cover, storage, designer, pine, finish, unit, design, sofa, chair, shelf, cabinet, plastic, birch, veneer, wall, door, seat, lamp, set
Flickr30K	man, woman, people, shirt, girl, boy, men, dog, street, child, women, person, water, children, group of people, hat, background, beach, ball, sidewalk.
COCO	man, people, woman, person, group, table, street, tennis, train, dog
CC3M	person, actor, artist, player, premiere, football, woman, beach, game, girl

Table 4: **Most common adjectives by dataset** ordered by count from left to right.

Dataset	Most Common Adjectives
Cars	new, open, diagnostic, necessary, short, upper, main, idle, high, other, normal, low, electric, manual, same, negative, positive, active, foreign, hot, old
IKEA	white, solid, black, clear, easy, adjustable, new, red, available, high, limited, green, washable, removable, extra, natural, low, last, soft, other, different, full, good,
Flickr30K	young, white, black, blue, red, little, green, other, large, yellow, small, brown, older, several, asian, gray, old, many, blond, dark.
COCO	white, two, large, small, front, red, young, black, young,
CC3M	white, young, day, black, red, blue, old, new, night, green, first,front

2.5 Choice of evaluation metric

We chose the text-to-image and image-to-text retrieval task for two main reasons: This metric is directly aligned with the popular contrastive training objective used for most of V&L models (e.g. CLIP or FLAVA) and as such should be their strongest suit. We however show that even under this metric CLIP under performs on FETA expert tasks compared to its performance demonstrated for e.g. photos of common objects. This metric is also possible to compute when regarding the automatic annotation process. In our automatic process, little is known a priori about the data. The assumption that co-location of text and images is strongly correlated with the semantics is a relatively general assumption. As shown in our Results in section 4.3 and in Table 3 of the main paper, our choice of automatic metric is valid as it gives the same conclusions using the same models and baselines performance as evaluated on the manually annotated and curated data.



Car-Manuals

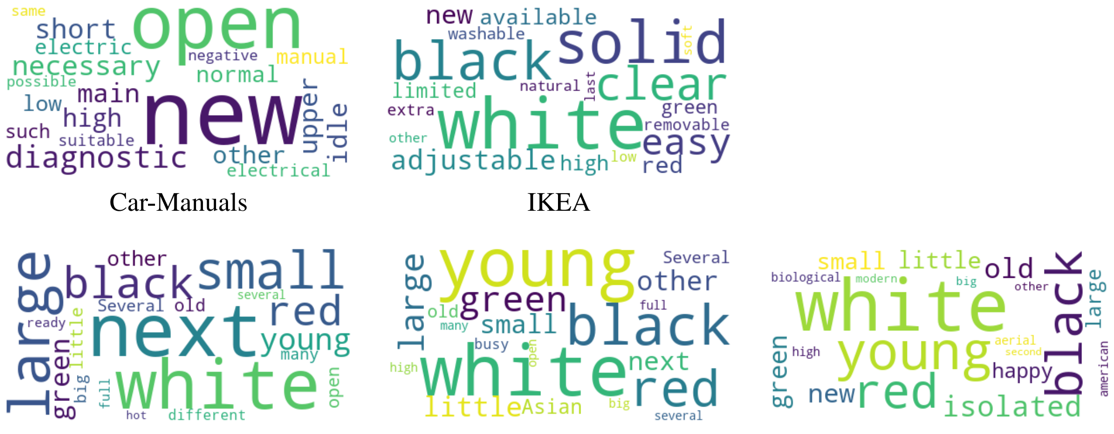
IKEA

COCO

Flickr30K

CC3M

Figure 2: A Visual representation of the most common nouns in the Car-Manuals and IKEA datasets, compared to COCO, Flickr30K, and CC3M



Car-Manuals

IKEA

COCO

Flickr30K

CC3M

Figure 3: A Visual representation of the most common adjectives in the Car-Manuals and IKEA datasets, compared to COCO, Flickr30K, and CC3M

3 Additional Implementation Details

Training Epochs: In general we train all MIL variants for 20 epochs, however we found that for the Zero-Shot and Few-Shot settings of the car manuals data we found that this amount results in over-fitting. We thus trained these two settings with only two epochs which we found to yield the best results.

4 Additional results

4.1 MIL Variants Ablation

In Section 3.2 of the paper we discussed several options for training CLIP under the MIL setting. Table 5 shows the performance for the three MIL variants on the Few-Shot and Many-Shot test settings. These results confirm that in the majority of the cases, MIL-NCE achieves the highest performing results. We therefore chose to use this MIL variant in all of experiments in the main paper where the MIL baselines are evaluated.

Table 5: **MIL Variant Ablations:** Image-to-Text and Text-to-Image retrieval accuracy for three MIL fine-tuning baseline variants under two different data-split settings. Our experimental settings is the same as Table 2 from the main paper but only include Many-Shot and Few-Shot (the more practical settings). The "Locked" column refers to versions trained with locked (frozen) parameters of the image encoder $\mathcal{M}_I$. Numbers in **bold** mark the best results while numbers in blue mark the second-best.

			Image-to-Text			Text-to-Image		
	Name	Locked	Rec@1	Rec@5	Rec@10	Rec@1	Rec@5	Rec@10
Few-Shot	CLIP-MIL-SOFTMAX		**14.6%**	34.4%	47.5%	13.4%	34.7%	47.8%
	CLIP-MIL-SOFTMAX	✓	13.4%	33.5%	46.6%	12.2%	34.1%	47.8%
	CLIP-MIL-MAX		13.7%	34.5%	47.8%	**15.7%**	**34.8%**	48.4%
	CLIP-MIL-MAX	✓	13.5%	33.9%	46.4%	12.3%	34.7%	48.5%
	CLIP-MIL-NCE		14.1%	**36.7%**	**48.9%**	15.0%	**35.2%**	**50.0%**
	CLIP-MIL-NCE	✓	13.8%	33.7%	47.5%	11.6%	33.0%	47.0%
Many-Shot	CLIP-MIL-SOFTMAX		32.0%	54.8%	65.7%	26.5%	58.1%	71.3%
	CLIP-MIL-SOFTMAX	✓	34.1%	56.5%	66.4%	26.7%	58.0%	70.3%
	CLIP-MIL-MAX		34.4%	56.7%	66.2%	27.3%	57.9%	71.1%
	CLIP-MIL-MAX	✓	31.2%	54.7%	65.7%	26.2%	58.1%	71.3%
	CLIP-MIL-NCE		32.6%	56.2%	**66.7%**	**27.8%**	**59.0%**	**72.3%**
	CLIP-MIL-NCE	✓	**34.5%**	**56.8%**	66.1%	27.2%	57.9%	70.7%

4.2 CLIP Architecture

In this section we present equivalent results to Table 2 from the main paper, using the ViT-L/14 backbone instead of ResNet50. Table 6 shows consistent baseline relative comparison results with the results presented in Table 2 of the main paper. As we can see from Table 6, using the ViT-L/14 backbone leads to higher performance (compared to ResNet50) with the CLIP-MIL options maintaining their significant gains under this stronger backbone.

4.3 Evaluating Additional Foundation Model

In Table2 of the main paper we have added the evaluation of three of the strongest performing and the most recent of the openly available foundation models - FLAVA [5], ALBEF[3] and VilT[2] to our set of Car-Manuals data evaluations comparing it to CLIP and our MIL baselines. For FLAVA, We used the released pre-trained FLAVA model available on Huggingface: https://huggingface.co/docs/transformers/model_doc/flava. We used facebook/flava-full pre-trained model. We evaluated both *contrastive* and *ITM* matching scores available in FLAVA and report results for the contrastive score, as it produced better results in all evaluations. We observe that stronger (relative to CLIP) common-objects performance reported by FLAVA, does not translate to improved numbers on FETA expert car manuals task. For ALBEF and VilT we have used the the VL-Checklist git repository [7]. We have used their code in order to obtain image-text scores between all image texts pairs in each document and used them in our retrieval test. As with FLAVA, the results for ALBEF and Vi;T are lower than CLIP and these models struggle in zero-shot performance on our data. We believe that this further supports our hypothesis that FMs need to be fine-tuned on the expert tasks as their out-of-the-box performance on these tasks is low. This is most likely due to bias towards common as is stated in the main paper. These experiments underline the need for our proposed FETA benchmark in order to improve the applicability of FMs to practical real-world problems often involving specialized expert V&L data.

4.4 Pre-trained CLIP vs Training from Scratch

All of the experiments presented in the main paper and the supplementary material initialize the training from a pre-trained CLIP model trained on 400M image-text pairs. In this section we examine the effect of training the ResNet50 CLIP model from scratch on the car manuals dataset. Table 7

Table 6: **Results using ViT-L/14 backbone** pre-trained by CLIP. Using the ViT-L/14 backbone leads to higher performance (compared to ResNet50) with the CLIP-MIL options maintaining their significant gains under this stronger backbone. Numbers in **bold** mark the best results while numbers in blue mark the second-best.

			Image-to-Text			Text-to-Image		
	Name	Locked	Rec@1	Rec@5	Rec@10	Rec@1	Rec@5	Rec@10
Zero-Shot	CLIP [4]		11.8%	29.5%	41.1%	12.0%	32.6%	46.6%
	Concatenate		12.0%	31.8%	44.6%	11.7%	31.7%	45.1%
	Concatenate	✓	12.2%	32.0%	44.6%	10.6%	30.8%	44.5%
	Choose-One		12.7%	31.4%	43.0%	11.8%	31.9%	46.0%
	Choose-One	✓	12.8%	31.6%	44.6%	11.2%	31.6%	45.3%
	CLIP-MIL		13.6%	32.9%	**46.5%**	**13.2%**	34.2%	**47.8%**
	CLIP-MIL	✓	**14.0%**	**33.5%**	46.3%	13.1%	**34.3%**	47.4%
One-Shot	CLIP [4]		11.8%	29.5%	41.1%	12.0%	32.6%	46.6%
	Concatenate		12.8%	24.4%	35.8%	11.4%	31.7%	45.9%
	Concatenate	✓	13.3%	33.8%	46.2%	11.2%	31.4%	45.2%
	Choose-One		12.6%	32.2%	44.4%	11.8%	32.3%	46.5%
	Choose-One	✓	13.2%	32.9%	44.6%	11.7%	32.4%	46.6%
	CLIP-MIL		14.2%	34.4%	46.8%	**13.7%**	**35.1%**	**49.2%**
	CLIP-MIL	✓	**14.5%**	**35.3%**	**47.2%**	13.0%	34.4%	48.1%
Few-Shot	CLIP [4]		10.7%	28.7%	40.6%	11.2%	30.8%	44.9%
	Concatenate		10.5%	31.2%	44.7%	11.9%	30.9%	46.9%
	Concatenate	✓	12.0%	31.5%	44.9%	13.0%	34.1%	48.7%
	Choose-One		14.7%	37.7%	51.7%	13.8%	36.3%	51.3%
	Choose-One	✓	12.9%	32.1%	45.1%	11.9%	33.5%	47.5%
	CLIP-MIL		**19.1%**	**44.2%**	**56.8%**	15.8%	**41.0%**	**56.1%**
	CLIP-MIL	✓	**19.1%**	40.7%	53.3%	**17.3%**	40.3%	53.8%
Many-Shot	CLIP [4]		15.7%	32.8%	43.2%	15.5%	39.7%	53.5%
	Concatenate		15.1%	32.1%	44.7%	14.3%	39.4%	54.8%
	Concatenate	✓	20.8%	37.6%	49.3%	19.0%	45.6%	60.3%
	Choose-One		24.1%	49.1%	61.1%	21.9%	53.3%	68.4%
	Choose-One	✓	30.9%	57.0%	67.9%	24.4%	57.3%	71.5%
	CLIP-MIL		31.2%	55.2%	66.2%	27.3%	60.1%	73.3%
	CLIP-MIL	✓	**40.1%**	**61.9%**	**71.0%**	**33.6%**	**65.1%**	**76.8%**

shows the result and compares the same model trained from scratch vs starting from a pre-trained CLIP model. We also include the locked and unlocked versions of the visual encoder. The advantage of starting from a strong pre-trained model is clear and raises the assumption that starting from a better pre-trained model will yield better results, also verified by the experiments using ViT-L/14 CLIP architecture in section 4.2.

4.5 IKEA US yearly catalogues

The IKEA data represents a different expert task - one of large-scale sales inventory (thousands of items). As with the technical documentation, sales catalogues naturally populate the long-tail of the common-objects biased data distributions used to train foundation models. Table 8 presents the results for the proposed baselines trained and tested on the IKEA dataset. Since in this case there are no distinct manufacturers and we did not like to partition on different yearly fashion styles due to their inconsistent nature, we followed a simple 5-fold cross-validation protocol using the entire IKEA data. Notably the IKEA dataset was processed using the same pipeline as the car manuals verifying the scalability of the proposed automatic annotation approach. As for the car manuals, also on IKEA data MIL based baselines obtaining significant advantages over other baselines. Interestingly, on IKEA data we also observe that both strategies that avoid direct use of the multiple annotation hypotheses (Concatenate and Choose-One) not only under-perform the MIL baselines, but also worsen the results

Table 7: **Initializing with a Pre-trained CLIP400M VS from Scratch**. Numbers in **bold** mark the best results while numbers in blue mark the second-best.

				Image-to-Text			Text-to-Image		
	Name	Pre-traind	Locked	Rec@1	Rec@5	Rec@10	Rec@1	Rec@5	Rec@10
Zero-Shot	CLIP-MIL	✓		10.5%	**34.0%**	**48.5%**	**11.7%**	**32.9%**	**47.9%**
		✓	✓	**11.0%**	29.2%	40.0%	9.7%	28.1%	40.6%
				3.4%	13.5%	23.0%	3.2%	16.3%	30.1%
			✓	4.2%	14.3%	24.6%	3.9%	18.3%	31.0%
One-Shot	CLIP-MIL	✓		11.0%	**30.3%**	**43.2%**	9.9%	27.9%	40.9%
		✓	✓	**11.9%**	**30.3%**	42.5%	**10.9%**	**29.4%**	**43.2%**
				5.2%	16.9%	28.1%	4.4%	19.2%	32.9%
			✓	5.3%	16.9%	27.4%	4.3%	17.8%	31.0%
Few-Shot	CLIP-MIL	✓		**14.1%**	**36.7%**	**48.9%**	**15.0%**	**35.2%**	**50.0%**
		✓	✓	13.8%	33.7%	47.5%	11.6%	33.0%	47.0%
				8.4%	25.1%	37.8%	6.7%	25.0%	39.0%
			✓	8.5%	24.9%	36.0%	6.1%	23.4%	36.6%
Many-Shot	CLIP-MIL	✓		32.6%	56.2%	**66.7%**	**27.8%**	**59.0%**	**72.3%**
		✓	✓	**34.5%**	**56.8%**	66.1%	27.2%	57.9%	70.7%
				29.1%	48.6%	58.1%	25.0%	52.3%	65.2%
			✓	24.7%	44.5%	54.7%	19.3%	46.6%	61.4%

relative to the CLIP baseline. This is likely due to sharper distinction between different associated text hypotheses, with only one of them being correct and other not only unrelated, but even belonging to other objects on the same (commonly densely packed) page. In such situation, non-MIL solutions are in significant disadvantage since as opposed to MIL they do not consider all the options at once with the logically accurate OR relation between possible labels, thus leading the model astray with wrongful gradient updates using losses computed with incorrect labels.

Table 8: **Results on IKEA dataset** using 5-fold cross-validation protocol on the entire IKEA US early manuals data. MIL based baselines obtain significant advantages over other baselines. Concatenate and Choose-One worsen the results of the CLIP baseline, likely due to only one of text label hypothesis (from the automatically extracted co-located set) being correct while others belong to other objects on the same (densely packed) page. In such situation, non-MIL solutions are in significant disadvantage since as they do not consider all the options at once with the logically accurate OR relation (as does MIL), thus leading the model astray with wrongful gradient updates from losses computed with incorrect labels. Numbers in **bold** mark the best results while numbers in blue mark the second-best.

			Image-to-Text			Text-to-Image		
	Name	Locked	Rec@1	Rec@5	Rec@10	Rec@1	Rec@5	Rec@10
All-Data	CLIP [4]		22.9%	43.3%	54.2%	25.5%	46.8%	59.5%
	Concatenate		6.7%	13.7%	18.2%	13.2%	27.0%	35.9%
	Concatenate	✓	8.1%	15.6%	20.6%	14.0%	26.9%	35.3%
	Choose-One		15.1%	30.2%	38.5%	17.9%	36.2%	46.4%
	Choose-One	✓	14.1%	28.0%	35.3%	16.4%	32.3%	41.8%
	CLIP-MIL		**26.8%**	**47.7%**	**57.8%**	**30.1%**	**54.4%**	**66.2%**
	CLIP-MIL	✓	24.4%	44.4%	54.7%	27.0%	49.9%	60.5%

4.6 Single fold reference results for quicker evaluation by future benchmark users

All experiments in the paper and other parts of this supplementary were performed using 5-fold cross-validation. Yet it is time consuming to evaluate many models 5 times. Therefore, as a service to the future users of our proposed FETA benchmark, in Table 9 we also provide reference results for a single fold out of the 5 we defined for the full evaluation. As can be seen from the table, the relative performance trends are preserved also in this single fold evaluation and hence it can serve as a reference for quicker evaluation for future users, before they run the full evaluation that takes 5 times longer. We provide a script to run this exact fold split in our code package for reproducibility.

Table 9: **Single fold reference results for quicker evaluation by future benchmark users**. Provided as a service to the future users of our proposed FETA benchmark by stating a 5 times faster (then full 5-fold cross val.) to compute reproducible evaluation reference point. Intended to facilitate faster evaluation & debugging of new methods. The exact split is enclosed in the benchmark code. Numbers in **bold** mark the best results while numbers in blue mark the second-best.

			Image-to-Text			Text-to-Image		
	Name	Locked	Rec@1	Rec@5	Rec@10	Rec@1	Rec@5	Rec@10
Zero-Shot	CLIP [4]		9.7%	26.6%	38.1%	10.1%	26.7%	39.4%
	Concatenate		6.5%	20.4%	31.5%	7.1%	25.0%	38.4%
	Concatenate	✓	9.4%	25.0%	36.7%	8.1%	24.0%	37.6%
	Choose-One		10.7%	27.6%	39.9%	9.3%	28.1%	41.8%
	Choose-One	✓	10.4%	26.7%	39.3%	9.2%	25.6%	37.9%
	CLIP-MIL		10.5%	**34.0%**	**48.5%**	**11.7%**	**32.9%**	**47.9%**
	CLIP-MIL	✓	**11.0%**	29.2%	40.0%	9.7%	28.1%	40.6%
One-Shot	CLIP [4]		9.7%	26.6%	38.1%	**10.1%**	26.7%	39.4%
	Concatenate		7.3%	22.0%	33.5%	7.6%	24.7%	39.6%
	Concatenate	✓	7.1%	20.5%	32.8%	6.9%	23.6%	37.9%
	Choose-One		8.9%	25.9%	37.8%	8.7%	28.2%	42.1%
	Choose-One	✓	7.5%	23.1%	35.3%	7.4%	23.9%	39.2%
	CLIP-MIL		**11.0%**	**28.7%**	**40.4%**	**10.1%**	**29.7%**	**44.1%**
	CLIP-MIL	✓	9.2%	24.8%	36.9%	8.1%	27.2%	41.6%
Few-Shot	CLIP [4]		8.6%	25.6%	37.2%	9.2%	24.3%	36.6%
	Concatenate		6.9%	24.8%	38.2%	10.7%	29.6%	45.4%
	Concatenate	✓	9.6%	23.0%	35.0%	10.3%	30.1%	42.1%
	Choose-One		11.7%	33.3%	45.3%	11.8%	33.0%	49.2%
	Choose-One	✓	14.9%	33.8%	47.6%	11.9%	29.2%	44.5%
	CLIP-MIL		**17.6%**	**40.8%**	**53.5%**	**16.0%**	**39.6%**	**53.9%**
	CLIP-MIL	✓	13.1%	37.1%	47.2%	12.0%	33.4%	48.3%
Many-Shot	CLIP [4]		13.8%	31.2%	41.6%	13.6%	36.4%	50.7%
	Concatenate		18.9%	38.1%	48.0%	16.9%	44.7%	59.9%
	Concatenate	✓	19.2%	38.0%	49.2%	15.8%	40.3%	54.9%
	Choose-One		22.5%	48.6%	60.5%	20.5%	50.9%	65.3%
	Choose-One	✓	28.1%	52.2%	63.4%	22.2%	52.9%	67.0%
	CLIP-MIL		31.7%	55.5%	66.6%	27.8%	**59.9%**	**72.6%**
	CLIP-MIL	✓	**35.5%**	**58.3%**	**67.0%**	**29.0%**	59.1%	71.6%

4.7 Median Results of Main Table 2

For completion we add the results of Few-Shot and Many-Shot settings from table 2 of the main manuscript with the difference of using Median instead of average

4.8 Illustrative explanation of our MIL variants

In Figure 4 we give an illustrative explanation of our MIL methods. Our MIL variant is composed of many to many MIL, many images to many texts, for the sake of simplicity we show here one to many MIL, image to texts MIL. Text to images MIL is very similar and thus not displayed. Each MIL block in the figure receives as inputs one image and many texts, a bag of positive texts chosen by our automatic annotation process and many negative texts taken from other pages the dataset. The output is detailed in Figure 4 and is used for the calculation of the loss. The figure is a schematic

Table 10: **Median results of Table 2 of the manuscript** Image-to-Text and Text-to-Image retrieval accuracy, same models and settings as in Table 2 but using median instead of average. Numbers in **bold** mark the best results while numbers in blue mark the second-best.

			Image-to-Text			Text-to-Image		
	Name	Locked	Rec@1	Rec@5	Rec@10	Rec@1	Rec@5	Rec@10
Few-Shot	CLIP [4]		9.1%	25.5%	36.2%	9.9%	23.7%	35.6%
	Concatenate		8.5%	24.2%	39.8%	10.2%	29.8%	45.1%
	Concatenate	✓	8.4%	23.3%	37.2%	10.0%	27.8%	40.6%
	Choose-One		11.1%	29.9%	44.3%	12.1%	33.4%	46.2%
	Choose-One	✓	11.7%	31.2%	44.2%	10.0%	29.6%	43.7%
	CLIP-MIL		**11.9%**	**33.9%**	**50.9%**	**14.3%**	**35.3%**	**53.1%**
	CLIP-MIL	✓	10.6%	31.3%	46.9%	9.9%	30.9%	47.4%
Many-Shot	CLIP [4]		13.8%	31.2%	41.6%	16.6%	36.4%	50.7%
	Concatenate		18.4%	38.1%	49.3%	16.1%	43.4%	59.6%
	Concatenate	✓	20.2%	40.1%	51.2%	16.5%	41.8%	59.9%
	Choose-One		24.6%	50.9%	62.9%	21.1%	53.0%	67.6%
	Choose-One	✓	28.0%	53.3%	64.8%	22.2%	52.9%	67.1%
	CLIP-MIL		31.8%	56.0%	**66.7%**	**27.9%**	**59.4%**	**72.3%**
	CLIP-MIL	✓	**34.3%**	**56.3%**	66.1%	27.3%	58.4%	71.2%

flowchart created for the purpose of intuition, the exact specifications of those losses are explained in Section 3 of the main paper.

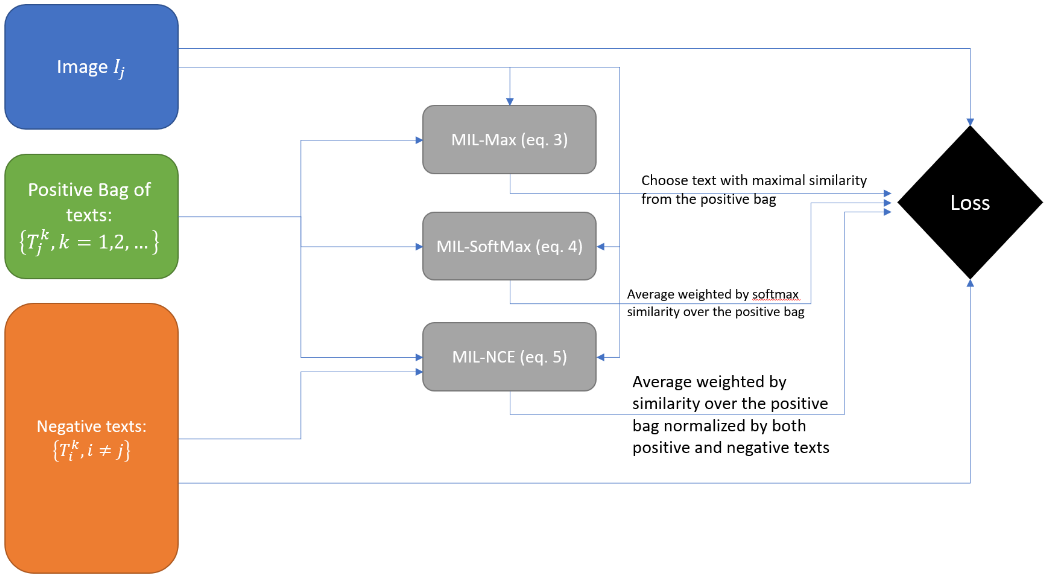


Image I_j
Positive Bag of texts: $\{T_j^k, k = 1,2, \dots\}$
Negative texts: $\{T_i^k, i \neq j\}$
MIL-Max (eq. 3)
MIL-SoftMax (eq. 4)
MIL-NCE (eq. 5)
Choose text with maximal similarity from the positive bag
Average weighted by softmax similarity over the positive bag
Average weighted by similarity over the positive bag normalized by both positive and negative texts
Loss

Figure 4: **Illustrative explanation of our MIL variants**: An example of one way MIL, from one image to a bag of texts. The figure is added for intuition only, the accurate details are in main paper section 3.

5 Code

The code is attached to this supplementary in "code.zip". The instructions for:

1. Automatic PDFs processing pipeline
2. Subsequent data pre-processing

3. Running the baselines
4. Evaluation of the baselines performance

are enclosed within in a contained README.md file. The APIs for the automatic PDF processing pipeline will become openly available upon acceptance. For the benchmark release we will also modify the massive runs scripts to support the main cluster configurations (e.g. SLURM), currently they are provided only for reference and utilize our internal (LSF) cluster architecture.

6 Data download instructions

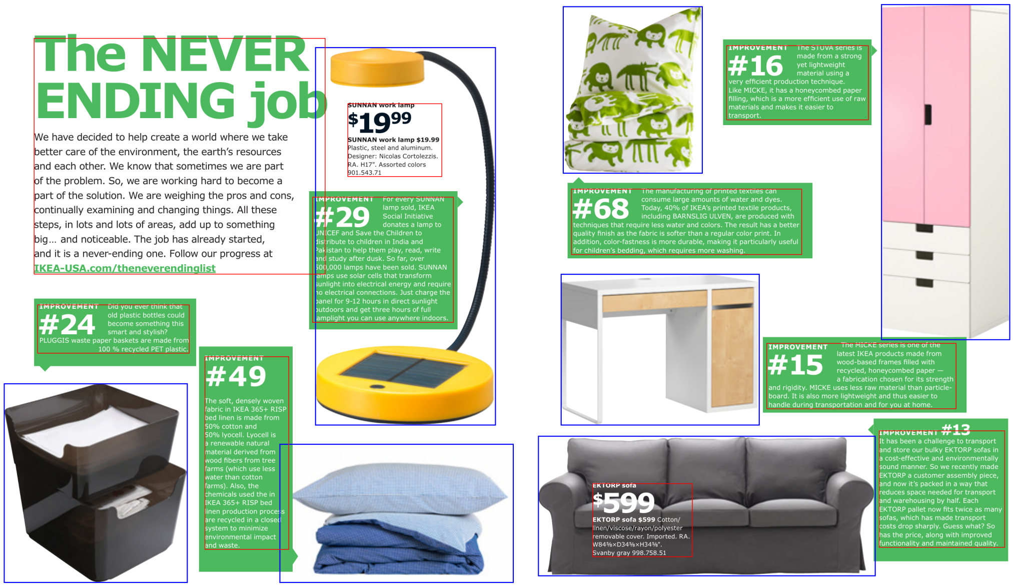
The data set is hosted on the IBM Cloud. In order to download the file please follow this commands:

```
» wget https://ai-vision-public-datasets.s3.eu.cloud-object-storage.appdomain.cloud/FETA/feta.tar.gz
» tar -xzvg feta.tar.gz
```

Inside the tar file there is README_data.md which explains how to use the data and how to easily obtain texts and images per document. We also add instructions here: In the tar, Car manuals and IKEA data are provided. To use the data: 1.Copy data to any desired path. 2. When running FETA code, set –train-data and –val-data to the pkl file inside each of these two data repositories. Detailed annotation and text of the entire data, pages and images are available in the pkl file which is located in the main directory of each dataset. We also provide humanly readable tsv files to make it easier to manually look at parts of the data. Each document has its own tsv file under <data_name>/texts/<doc_name>.tsv . Inside each tsv file, which can be opened in Excel or as a text file, there are four columns listing the texts and main annotations of the document: page_number, text_ind, text, bbox.

7 License

This dataset is freely available: it can be redistribute under the terms of the GNU General Public License (version 3) as published by the Free Software Foundation. The full license is attached to the supplementary material as License.md



The NEVER ENDING job

We have decided to help create a world where we take better care of the environment, the earth's resources and each other. We know that sometimes we are part of the problem. So, we are working hard to become a part of the solution. We are weighing the pros and cons, continually examining and changing things. All these steps, in lots and lots of areas, add up to something big… and noticeable. The job has already started, and it is a never-ending one. Follow our progress at IKEA-USA.com/theneverendinglist

SUNNAN work lamp
$19.99
SUNNAN work lamp $19.99 Plastic, steel and aluminum. Designer: Nicolas Cortolezzis. RA. H17". Assorted colors 901.543.71

IMPROVEMENT #29 For every SUNNAN lamp sold, IKEA Social Initiative donates a lamp to UNICEF and Save the Children to distribute to children in India and Pakistan to help them play, read, write and study after dusk. So far, over 500,000 lamps have been sold. SUNNAN lamps use solar cells that transform sunlight into electrical energy and require no electrical connections. Just charge the panel for 9-12 hours in direct sunlight outdoors and get three hours of full lamplight you can use anywhere indoors.

IMPROVEMENT #24 Did you ever think that old plastic bottles could become something this smart and stylish? PLUGGIS waste paper baskets are made from 100 % recycled PET plastic.

IMPROVEMENT #49 The soft, densely woven fabric in IKEA 365+ RISP bed linen is made from 50% cotton and 50% lyocell. Lyocell is a renewable natural material derived from wood fibers from tree farms (which use less water than cotton farms). Also, the chemicals used the in IKEA 365+ RISP bed linen production process are recycled in a closed system to minimize environmental impact and waste.

IMPROVEMENT #16 The STUVA series is made from a strong yet lightweight material using a very efficient production technique. Like MICKE, it has a honeycombed paper filling, which is a more efficient use of raw materials and makes it easier to transport.

IMPROVEMENT #68 The manufacturing of printed textiles can consume large amounts of water and dyes. Today, 40% of IKEA's printed textile products, including BARNSLIG ULVEN, are produced with techniques that require less water and colors. The result has a better quality finish as the fabric is softer than a regular color print. In addition, color-fastness is more durable, making it particularly useful for children's bedding, which requires more washing.

IMPROVEMENT #15 The MICKE series is one of the latest IKEA products made from wood-based frames filled with recycled, honeycombed paper — a fabrication chosen for its strength and rigidity. MICKE uses less raw material than particleboard. It is also more lightweight and thus easier to handle during transportation and for you at home.

EKTORP sofa
$599
EKTORP sofa $599 Cotton/linen/viscose/rayon/polyester removable cover. Imported. RA. W84⅝×D34⅝×H34⅝". Svanby gray 998.758.51

IMPROVEMENT #13 It has been a challenge to transport and store our bulky EKTORP sofas in a cost-effective and environmentally sound manner. So we recently made EKTORP a customer assembly piece, and now it's packed in a way that reduces space needed for transport and warehousing by half. Each EKTORP pallet now fits twice as many sofas, which has made transport costs drop sharply. Guess what? So has the price, along with improved functionality and maintained quality.

Figure 5: **Examples from the IKEA dataset**: In this example we can see the bounding box of all the detected texts withing the page marked in a red box. The images are marked in a blue box.

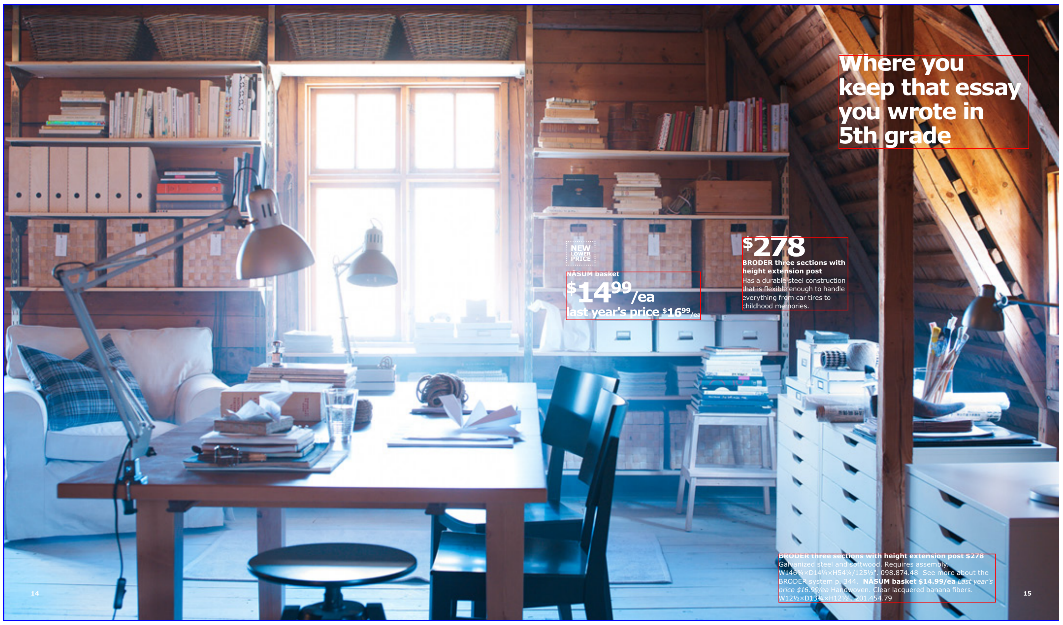

Where you keep that essay you wrote in 5th grade
NEW LOWER PRICE
NÄSUM basket
$14.99/ea
last year's price $16.99/ea
$278
BRODER three sections with height extension post
Has a durable steel construction that is flexible enough to handle everything from car tires to childhood memories.
14
15

Figure 6: **Examples from the IKEA dataset**: In this example we can see the bounding box of all the detected texts withing the page marked in a red box.The images are marked in a blue box.

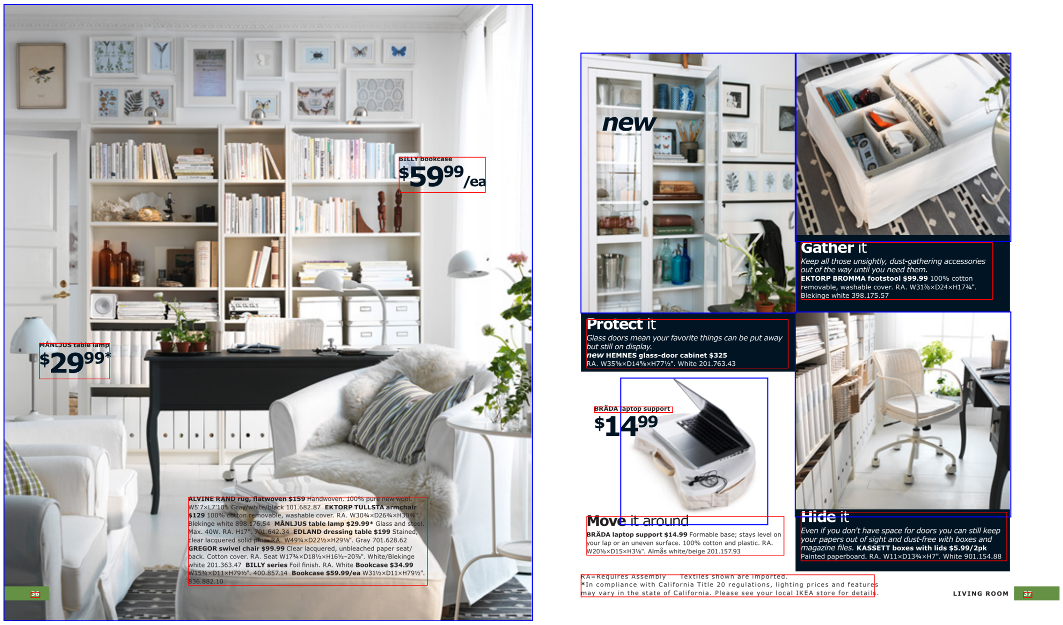
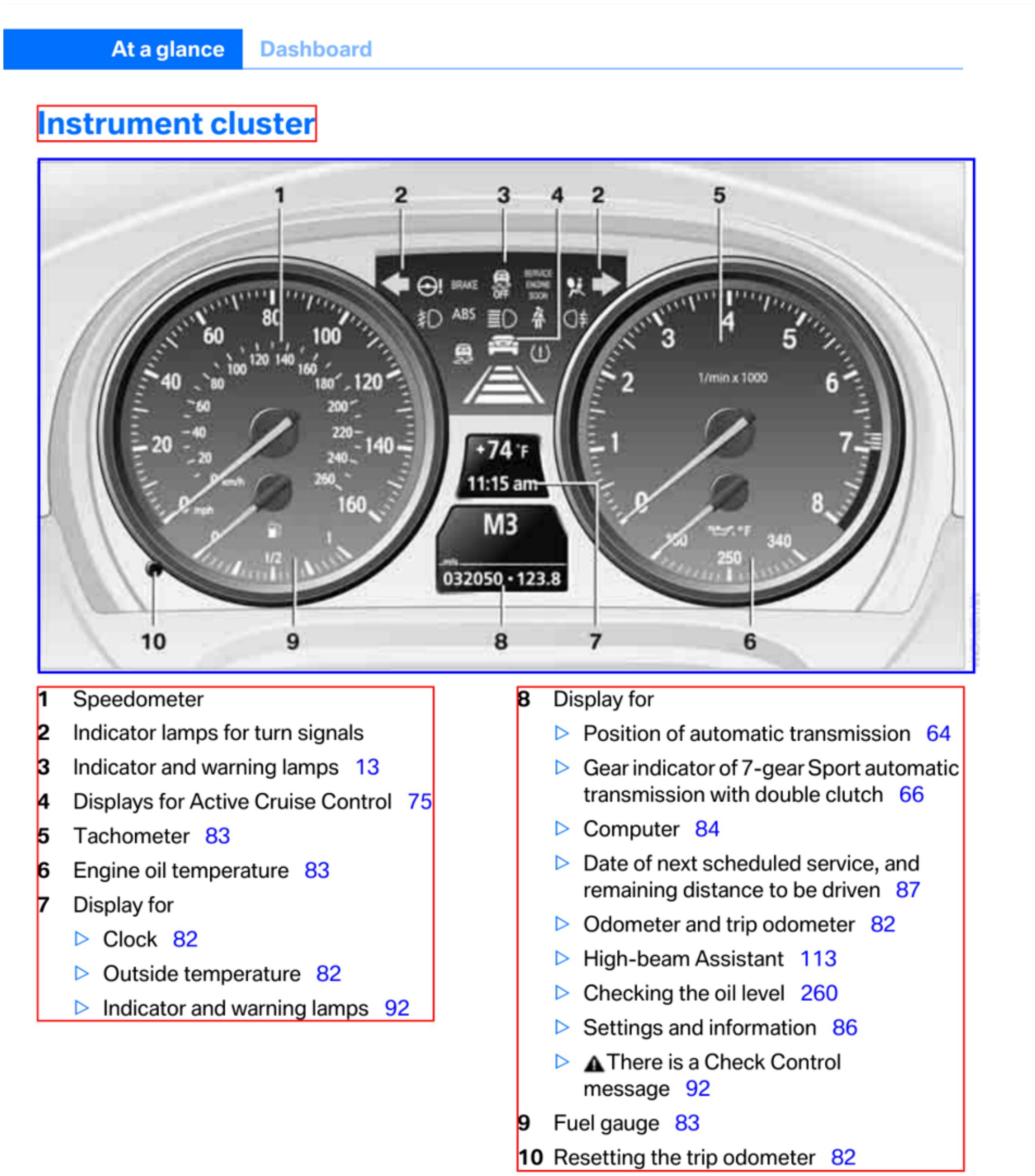

BILLY bookcase
$59.99/ea
MÅNLJUS table lamp
$29.99*
new
Gather it
Keep all those unsightly, dust-gathering accessories out of the way until you need them.
Protect it
Glass doors mean your favorite things can be put away but still on display.
BRÅDA laptop support
$14.99
Move it around
Hide it
Even if you don't have space for doors you can still keep your papers out of sight and out-of-mind with boxes and magazine files.
LIVING ROOM

Figure 7: **Examples from the IKEA dataset**: In this example we can see the bounding box of all the detected texts withing the page marked in a red box. OnlyThe images are marked in a blue box.



At a glance Dashboard

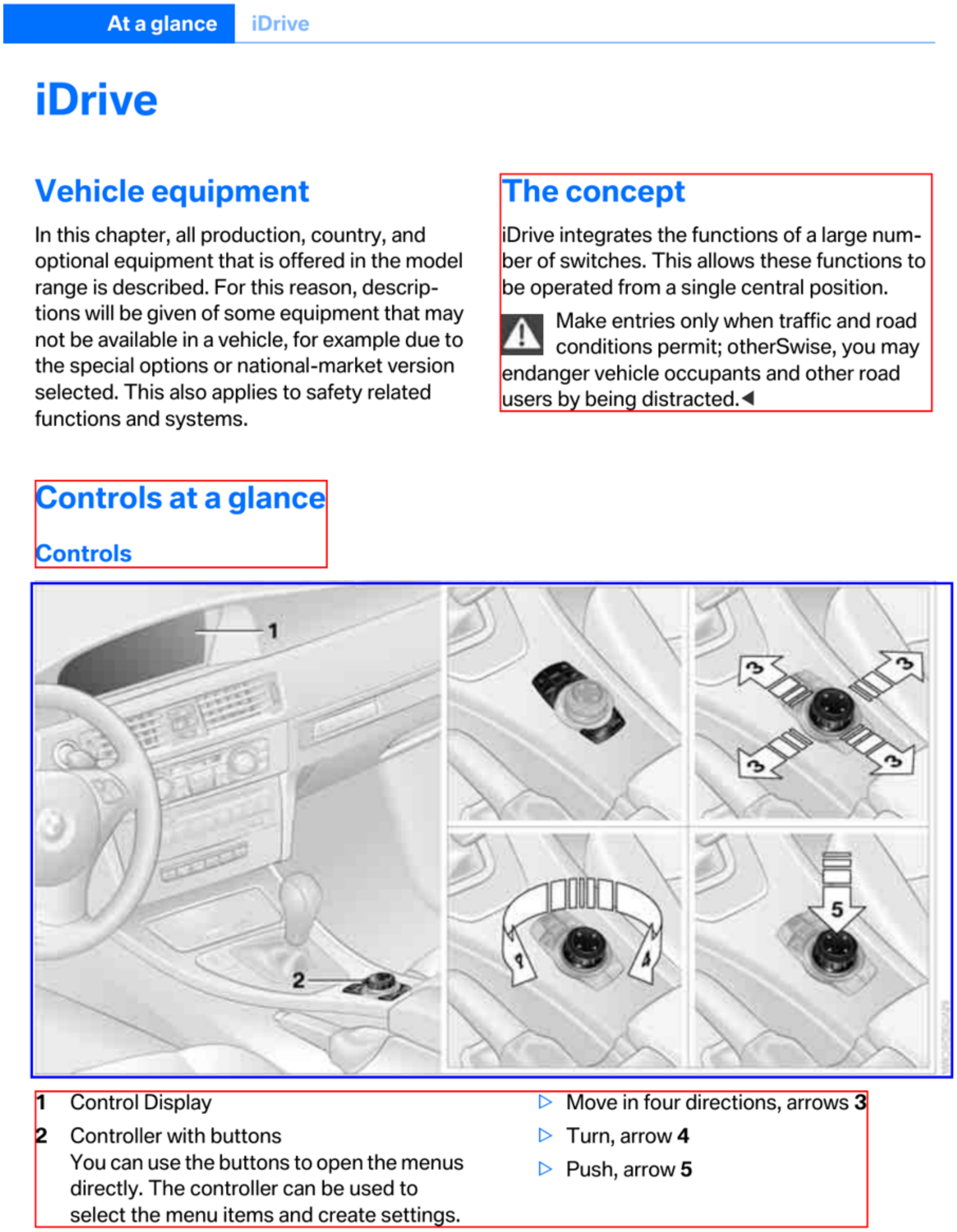
Instrument cluster

1 Speedometer
2 Indicator lamps for turn signals
3 Indicator and warning lamps 13
4 Displays for Active Cruise Control 75
5 Tachometer 83
6 Engine oil temperature 83
7 Display for
- ▷ Clock 82
- ▷ Outside temperature 82
- ▷ Indicator and warning lamps 92

8 Display for
- ▷ Position of automatic transmission 64
- ▷ Gear indicator of 7-gear Sport automatic transmission with double clutch 66
- ▷ Computer 84
- ▷ Date of next scheduled service, and remaining distance to be driven 87
- ▷ Odometer and trip odometer 82
- ▷ High-beam Assistant 113
- ▷ Checking the oil level 260
- ▷ Settings and information 86
- ▷ ⚠There is a Check Control message 92

9 Fuel gauge 83
10 Resetting the trip odometer 82

Figure 8: **Examples from the expert car manuals dataset**: In this example we can see the bounding box of all the detected texts withing the page marked in a red box. The images are marked in a blue box.



At a glance iDrive

iDrive

Vehicle equipment

In this chapter, all production, country, and optional equipment that is offered in the model range is described. For this reason, descriptions will be given of some equipment that may not be available in a vehicle, for example due to the special options or national-market version selected. This also applies to safety related functions and systems.

The concept

iDrive integrates the functions of a large number of switches. This allows these functions to be operated from a single central position.

⚠ Make entries only when traffic and road conditions permit; otherSwise, you may endanger vehicle occupants and other road users by being distracted.◀

Controls at a glance

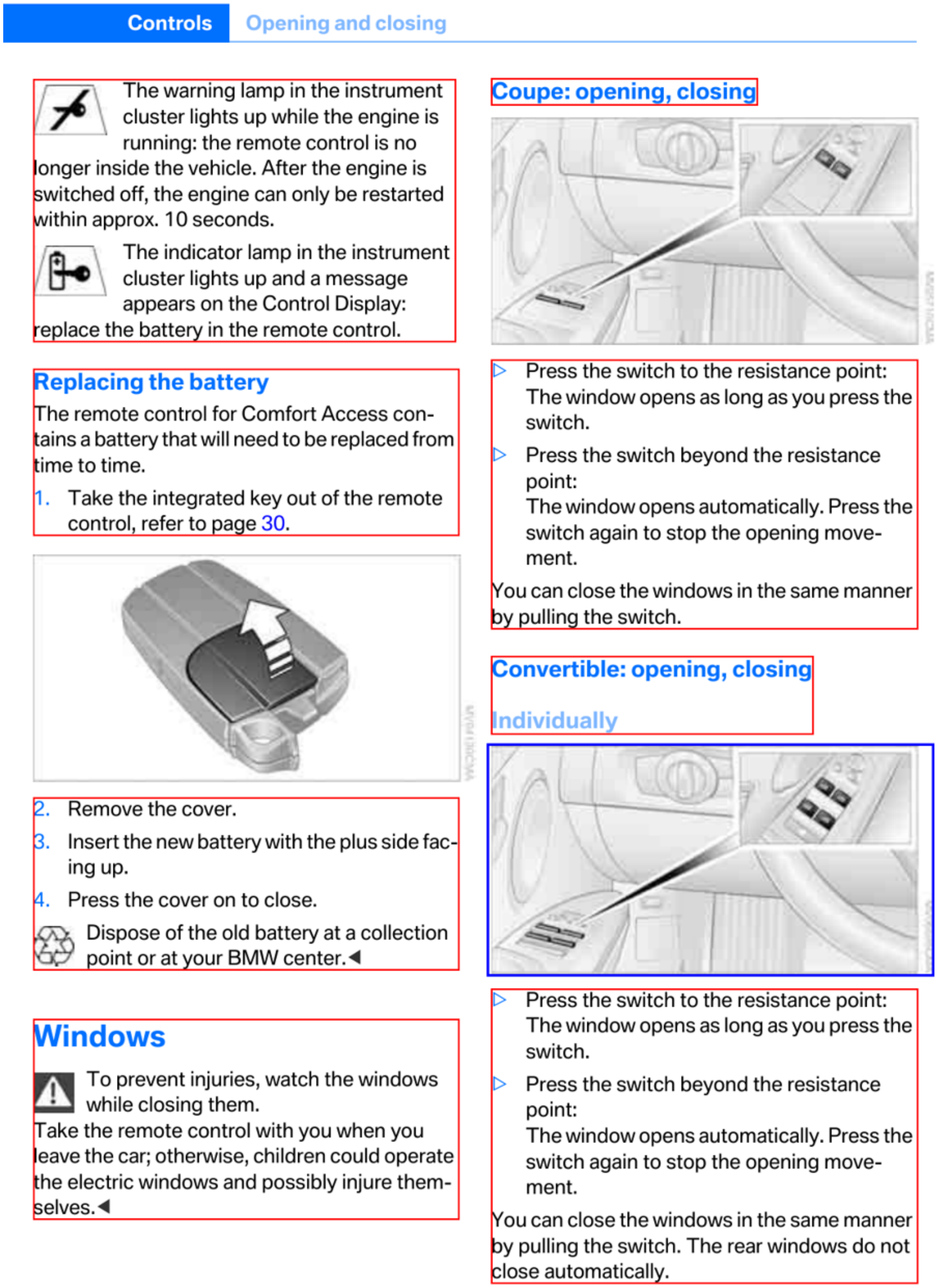
Controls

1 Control Display

2 Controller with buttons
You can use the buttons to open the menus directly. The controller can be used to select the menu items and create settings.

- ▷ Move in four directions, arrows **3**
- ▷ Turn, arrow **4**
- ▷ Push, arrow **5**

Figure 9: **Examples from the expert car manuals dataset**: In this example we can see the bounding box of all the detected and **associated** texts withing the page marked in a red box. The images are marked in a blue box.



Controls Opening and closing

The warning lamp in the instrument cluster lights up while the engine is running: the remote control is no longer inside the vehicle. After the engine is switched off, the engine can only be restarted within approx. 10 seconds.

The indicator lamp in the instrument cluster lights up and a message appears on the Control Display: replace the battery in the remote control.

Replacing the battery

The remote control for Comfort Access contains a battery that will need to be replaced from time to time.

1. Take the integrated key out of the remote control, refer to page 30.
2. Remove the cover.
3. Insert the new battery with the plus side facing up.
4. Press the cover on to close.

Dispose of the old battery at a collection point or at your BMW center.◀

Windows

To prevent injuries, watch the windows while closing them.
Take the remote control with you when you leave the car; otherwise, children could operate the electric windows and possibly injure themselves.◀

Coupe: opening, closing

- ▷ Press the switch to the resistance point: The window opens as long as you press the switch.
- ▷ Press the switch beyond the resistance point: The window opens automatically. Press the switch again to stop the opening movement.

You can close the windows in the same manner by pulling the switch.

Convertible: opening, closing

Individually

- ▷ Press the switch to the resistance point: The window opens as long as you press the switch.
- ▷ Press the switch beyond the resistance point: The window opens automatically. Press the switch again to stop the opening movement.

You can close the windows in the same manner by pulling the switch. The rear windows do not close automatically.

Figure 10: **Examples from the expert car manuals dataset**: In this example we can see the bounding box of all the detected and **associated** texts withing the page marked in a red box. One image is marked in a blue box.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1

[2] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021. 6

[3] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation, 2021. 6

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 7, 8, 9, 10

[5] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model, 2021. 1, 6

[6] Jae-Chern Yoo and Tae Hee Han. Fast normalized cross-correlation. *Circuits, systems and signal processing*, 28(6):819–843, 2009. 2

[7] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations, 2022. 6

[8] Yi Zheng, Qitong Wang, and Margrit Betke. Semantic-based sentence recognition in images using bimodal deep learning. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2753–2757, 2021. 1