A. Data scaling analysis

To analyze the influence of data amount, we train MatCha using different amount (25%, 50%, 75%, 100%) of our LaMenDa data, and report results in the Tab. 6. We use the same setting and data as row-5 in Tab-2, except that the training schedule is shorter (5k iterations) due to time limitations. As shown, improvements of our generated data follow a typical data scaling law: improvements are large initially (at 25%) then continues with smaller margins.

data amount	**0%**	**25%**	**50%**	**75%**	**100%**
human	30.21	34.38	35.73	36.25	37.19
augment	65.31	66.87	68.85	69.69	68.96
avg.	47.76	50.63	52.29	52.97	53.07

Table 6. Results with different amount of generated data.

B. Template-based QA generation

Results using template-generated data only. We report the result using template-generated QA data only, under the same setting as Tab-2 (1024 patches, on ChartQA): training with only template QAs leads to 54.32% strict accuracy (38.23% on human, 70.73% on augmented). This result is better than baseline (47.76%), but lower than LLM generator which gets 58.65%. Additionally, although template QAs are very clean, they (a) are not free-form and (b) rely on groundtruth SVG metadata, thus cannot generalize to images without SVG, which highlights the advantages of LLM generator.

Template list. Empirically, we find that template diversity is crucial for the model performance, otherwise the LLM generator may overfit to rigid templates. For example, in the early stage of the project, our LLM generator gets only 49.8% accuracy trained with 16 templates without rephrasing, which is much lower than the current 55.2% (row 5 Tab-2). Tab. 7 shows the final list of templates that are used for the templated-based question-answer generation pipeline. For each template, we manually define a functional program as inspired by the CLEVR dataset [21]. The functions include 7 basic operations like `SUM`, `COUNT`, `COMPARE`, and a `VQA` operation. The execution is motivated by VisProg [16], where each operation is executed by a predefined Python function.

C. Prompts

Tab. 8 shows the prompts to prompt the LLM-based data generator, for controllably generating questions and answers.

D. Dataset statistics

Tab. 9 shows detailed statistics for the ChartQA and the PlotQA datasets. Tab. 10 shows the detailed statistics of the chart captioning datasets. We generated questions and answers using our LLM-based data generator for the chart images in these chart captioning datasets.

E. Additional results

Tab. 11 shows the full results including relaxed accuracy and strict accuracy for Tab. 4 in the main paper. Tab. 12 shows the full results including relaxed accuracy and strict accuracy for Tab. 5 in the main paper.

F. Examples comparing MatCha with and without LaMenDa

Color
1. What color is $\langle N \rangle$ represented?
2. What is the value of the $\langle C \rangle \langle F \rangle$?
3. Which category is represented by $\langle C \rangle$?
Spatial
4. What is the value of the $\langle S \rangle$ bar?
5. What does the $\langle S \rangle$ bar represent?
6. What is the value of the second bar from the $\langle S \rangle$?
7. What is [represented by] the second bar from the $\langle S \rangle$?
8. What is the value of the third bar from the $\langle S \rangle$?
9. What is represented by the third bar from the $\langle S \rangle$?
Count
10. How many $\langle F \rangle$s are shown in the plot?
11. How many $\langle C \rangle$ bars are shown in the plot?
12. How many colors are used to represent the $\langle F \rangle$s in the plot?
Math
13. What is the average of $\langle N \rangle$?
14. What is the max [value] of $\langle N \rangle$?
15. What is the min value of the $\langle N \rangle$?
16. What is the [total] sum [value] of $\langle L \rangle$ and $\langle L2 \rangle$?
17. What is the difference between [values of] $\langle L \rangle$ and $\langle L2 \rangle$?
18. What is the value of the smallest category[in the chart]?
19. What is the value of the largest category[in the chart]?
20. What is the smallest category?
21. What is the largest category?
22. What is the average [value] of the two smallest categories[in the chart]?
23. What is the average [value] of the two largest categories[in the chart]?
24. What is the difference between the largest [category] and the smallest category?
25. What is the ratio [value] of $\langle L \rangle$ and $\langle L2 \rangle$?
26. How many times [is] $\langle L \rangle$ bigger than $\langle L2 \rangle$?
27. What is the average of $\langle L \rangle$ and $\langle L2 \rangle$?
28. Is [the value of] $\langle L \rangle$ more than $\langle L2 \rangle$?

Table 7. List of templates.

1. The question should be similar to this: ...
2. The question should be free form.
3. The question should require color understanding of the image.
4. The question should require counting.
5. The question should require counting of colors.
6. The question should require counting and color understanding.
7. The question should require spatial understanding of the image.
8. The question should require math reasoning about min.
9. The question should require math reasoning to compute min.
10. The question should require math reasoning to compute average of two categories.
11. The question should require math reasoning to compute average.
12. The question should require math reasoning to compute max.
13. The question should require math reasoning about the difference between max and min.
14. The question should require math reasoning to compute difference.
15. The question should require math reasoning about comparison.
16. The question should require math reasoning about average and max.
17. The question should require math reasoning to compute sum.
18. The question should require math reasoning about max.
19. The question should require math reasoning about average and min.
20. The question should require math reasoning to compute ratio.
21. The question should require color understanding and math reasoning to compute difference.
22. The question should require color understanding and math reasoning about comparison.
23. The question should require spatial understanding and math reasoning to compute difference.
24. The question should require spatial understanding and math reasoning about average.

Table 8. List of prompts for prompting the LLM-based data generator.

	ChartQA			**PlotQA**		
	Images	**HumanQA**	**AugmentedQA**	**Images**	**V1 QA**	**V2 QA**
train	18,317	7,398	20,901	157,070	5,733,893	20,249,479
val	1,056	960	960	33,653	1,228,468	4,360,648
test	1,509	1,250	1,250	33,660	1,228,313	4,342,514

Table 9. Statistics for ChartQA and PlotQA.

		#images	#captions	#llava_pred	#filter-10
Chart-to-text [24]	statista_two_col	27868	27868	603283	613096
	statista_multi_col	6943	6943	152746	107752
	pew_two_col	1486	1486	31806	23521
	pew_multi_col	7799	7799	171578	113967
VisText [52]	vistext	8822	9969	172319	76788
ChartSumm [47]	img_list_s	40985	32786	901670	364218
	img_list_k	43378	34702	954316	336022
All	-	137281	121553	2987718	1635364

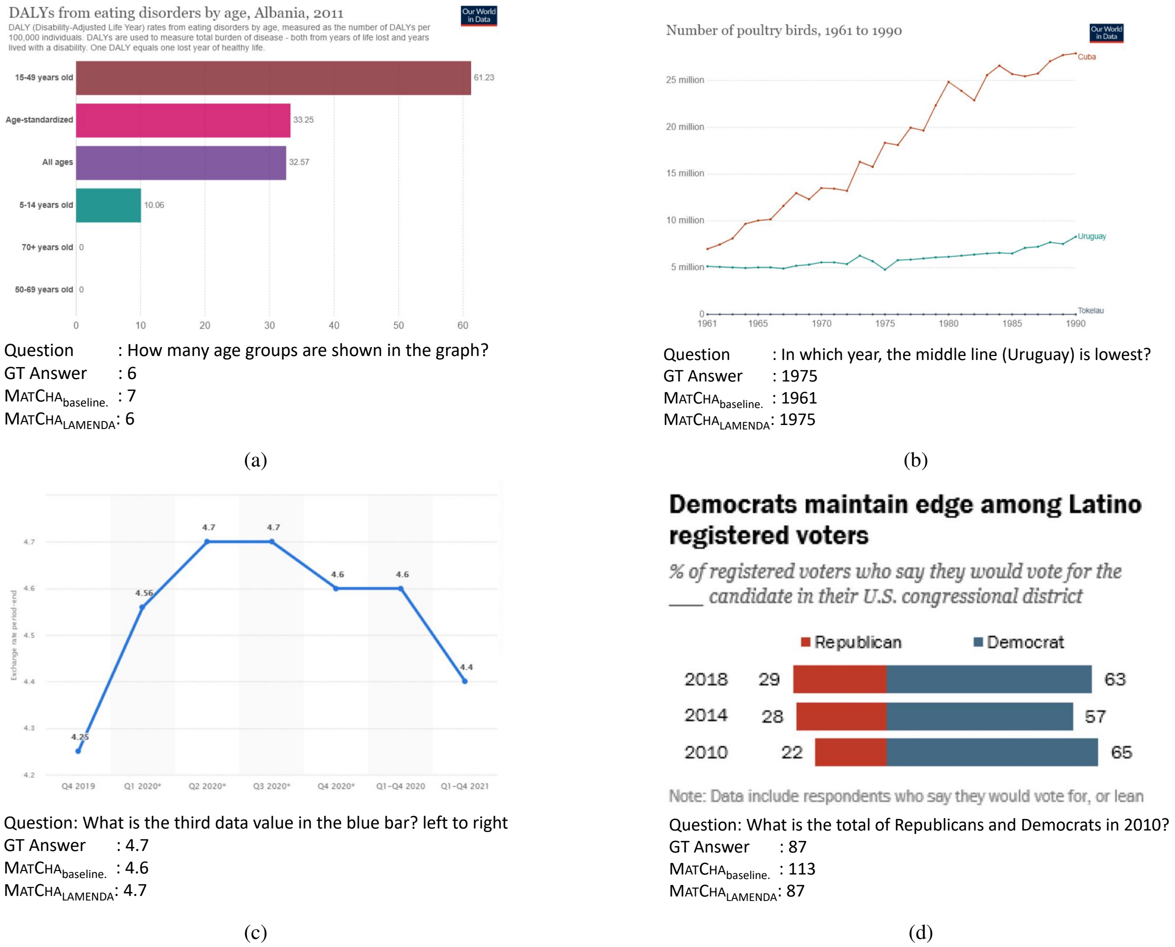
Table 10. Statistics for the chart captioning datasets.

	Accuracy			Relaxed Accuracy		
	avg	V1	V2	avg	V1	V2
VisionTapas-OCR [40]	-	-	-	53.90	65.30	42.50
VL-T5-OCR [40]	-	-	-	65.96	75.90	56.02
DePlot+FlanPaLM(540B)+Codex [34]	-	-	-	66.6	62.2	71.0
Pix2Struct [28]	-	-	-	72.5	73.2	71.9
MatCha [35]	-	-	-	**91.5**	**92.3**	**90.7**
MatCha$_{1024}$ (reimpl.)	41.58	66.66	16.50	74.95	73.88	76.02
MatCha$_{1024}$ + LaMenDa	43.04	68.48	17.60	78.41	78.44	78.38
MatCha$_{4096}$ (reimpl.)	50.89	76.14	25.64	91.97	92.64	91.30
MatCha$_{4096}$ + LaMenDa	**51.40**	**76.42**	**26.38**	**92.89**	**93.94**	**91.84**

Table 11. Full results for Tab. 4: comparison with SoTAs on PlotQA test split. With our generated data, MatCha achieves the SoTA performance.

	# Questions	Accuracy			Relaxed Accuracy		
		avg	human	augment	avg	human	augment
baseline	28,299	47.76	30.21	65.31	58.54	39.58	77.50
+colors	+46,512	47.81	29.90	65.73	59.32	40.21	78.44
+spatial	+49,387	49.27	30.63	67.92	61.09	43.33	78.85
+count	+36,705	47.40	29.17	65.63	58.28	38.54	78.02
+minmax	+71,788	49.58	31.87	67.29	59.38	40.63	78.13
+average	+50,717	48.96	31.35	66.56	60.00	40.52	79.48
+compare	+14,690	47.66	28.85	66.46	58.59	38.85	78.33
+calculation	+56,361	50.52	33.85	67.19	62.40	45.21	79.58
+all	+326,160	54.79	39.27	70.31	66.15	50.42	81.88

Table 12. Full results for Tab. 5: a detailed look at the effect for different question types. Strict accuracy on ChartQA val split is shown. Each type helps and combining all of them leads to a further performance gain.



Question : How many age groups are shown in the graph?
GT Answer : 6
MatCha$_{\text{baseline.}}$: 7
MatCha$_{\text{LAMENDA}}$: 6

(a)

Question : In which year, the middle line (Uruguay) is lowest?
GT Answer : 1975
MatCha$_{\text{baseline.}}$: 1961
MatCha$_{\text{LAMENDA}}$: 1975

(b)

Question: What is the third data value in the blue bar? left to right
GT Answer : 4.7
MatCha$_{\text{baseline.}}$: 4.6
MatCha$_{\text{LAMENDA}}$: 4.7

(c)

Question: What is the total of Republicans and Democrats in 2010?
GT Answer : 87
MatCha$_{\text{baseline.}}$: 113
MatCha$_{\text{LAMENDA}}$: 87

(d)

Figure 6. The baseline MatCha fails to correctly answer the question which needs visual understanding and then doing arithmetic operation, whereas MatCha fine-tuned with our data is able to correctly answer it.