Appendix of HiRes-LLaVA: Restoring Fragmentation Input in High-Resolution Large Vision-Language Models

A. Implementation Details

Training datasets. Table 1 shows the detailed dataset construction of the capability enhancement stage of HiRes-LLaVA. Specifically, it has 830K captioning including the ShareGPT4V [7], ShareGPT4o [16] and ALLAVA [5]. There are 821K OCR data from SynthDoG [15] including English OCR data as well as MMC-Alignment [25], UReader [41], K12 printed [1] which is a short OCR dataset. There is also 200K text instruction data from Magpie Pro [40], sampling from the data generated by Llama3.1-70B, Llama3-70B, and Qwen2-72B.

Task	Datasets(# Sample)	Sum
Caption	ShareGPT4V(89k), ALLAVA4V(684k), ShareGPT-4O(57k).	830K(44.8%)
OCR	SynthDoG-EN(300k), MMC-Alignment(200k), UReader(101k), K12 printed(120k), SynthDoG-ZH(100k).	821k(44.4%)
Text	Magpie Pro(200k)	200k(10.8%)
Total		**1.8M**

Table 1. Datasets in the capability enhancement stage.

Table 2 shows the detailed construction of the 3M instruction tuning dataset. First, we remove 23K caption data and ShareGPT data from original LLaVA-158K [26] and include GPT4V/GPT4o-generated caption data, i.e., LAION-GPT4v [17], ShareGPT4V [7], ShareGPT4o [16] and ALLAVA instruction data [5]. To enhance the common knowledge of our model, we convert the visual spatial reasoning [23], AI2D [13], and Science QA [29] training set into the instruct-tuning data. To activate the understanding science, we collect data from ViQuAE [19], TextbookQA [14], IconQA [28] and sampled 50k data from the Cambrian's Data Engine [38]. We also collect document-oriented data from diverse datasets, includes ChartQA [30], DVQA [12], PlotQA [32], OCRVQA [33], ST-VQA [3], DocVQA [10], InfoVQA [31], DeepForm [37], TAT-DQA [42], TabFact [8], LRV-Chart[24] and WebSRC [9]. We merge some datasets from Cauldron [18], including RAVEN, ROBUT-SQA, ROBUT-WTQ, HiTab, IAM, Rendered Text, ORAND-CAR-A, Visual7W, Chart2Text, AI2D, vistext, Diagram-image-to-text.

Module Design Details. The self-mining sampler consists of one cross-attention block with an output layer norm. The cross-attention block has a cross-attention layer and a FFN. Both of them apply the residual shortcut. The cross-attention layer has two layer norm for the query and key/value, respectively. As for the SliceRestore Adapter, the parameters of the self-attention layer with the layer norm are initialized from the pretrained CLIP self-attention at the same depth. To provide the positional information between slices, we apply a 2D RoPE [35, 36] on the global fusion module.

Training pipeline. We list the hyperparameters for the three-stage training at Tab. 3.

Evaluation details. We utilize the open-source evaluation tools, lmms-eval [20], to align our evaluation method to LLaVA-NeXT [27].

Benchmark construction. In our EntityGrid-QA, the construction of multiple choices is a vital part of EntityGrid-QA. For different types of entities, we apply different augmentations to obtain the other three choices for each question. For text and decimal, we randomly delete, add, or change one letter or digit. The object figures are collected from the COCO dataset [22]. For both categories of icons and objects, we use GPT-4 to list three other entities' names with similar appearance as the negative options.

B. More Ablation

Comparison on the Same Training Set To demonstrate the effectiveness of our method, we compare the performance of LLaVA-1.5 and our method trained on the same data. Specifically, we train these two models on two different scale training data set, *i.e.*, LLaVA-655K [21] and LLaVA-655K with additional Doc-79K data (the dataset of our ablation setting). Results from Tab. 4 show that adding 79K document data can highly improve models' performance on ChartQA, DocQA and InfoVQA but slightly drops the performance on MM-Bench and MME-Perception. Hires-LLaVA outperforms the

Task	Datasets(# Sample)	Sum
General QA	LLaVA(135K), ALLaVA(660K) VQAv2(83K), GQA(72K), OKVQA(9K), A-OKVQA(66K), VSR(12K), ShareGPT4V(89K), TextCaps(22K), Laion-GPT4V(11K), ShareGPT-4O(57K), RAVEN(3K), Visual7w(14K), RefCOCO(48K), VG(86K)	1.4M (48.0%)
Science	ScienceQA(19K), ai2d(14K), ViQuAE(4K), TextbookQA(21K), IconQA(30K), Data Engine(50K)	139K(4.6%)
Doc QA/OCR	OCRVQA(80K), TextVQA(57K), SynthDog(30K), LLaVAR(39K), WikiTableQuestions(29K), KleisterCharity(15K), iiit(6K), MLHME(30K), VisualMRC(19K), ChartQA(48K), DocVQA(102K), InfoVQA(33K), DVQA(200K), PlotQA(10K), TAT-DQA(2K), TableFact(65K), WebSRC(5K) DeepForm(8K), Chart2text(27K) Vistext(10K), chrome writting(9K), IAM(6K), Rendered text (10K), Orand-CAR-A(2K), lrv-chart(2K), ROBUT-SQA(9K), ROBUT-WTQ(4K), Hitab(3K), Diagram-image-to-text(0.3K).	0.9M(30.1%)
Code Generation	WebSight(50K)	50K(1.7%)
Text-only	Magpie-Pro(150K), Evol(142K), mathinstruct(81K), mathplus(95K).	469K(15.6%)
Total		**3M**

Table 2. Summary of datasets used in the instruction tuning stage.

	Settings	Stage-1	Stage-2	Stage-3
Vision	**Resolution**	$448\times\{\{1\times2\},\cdots,\{3\times3\}\}$	$448\times\{\{1\times2\},\cdots,\{3\times3\}\}$	$448\times\{\{1\times2\},\cdots,\{3\times3\}\}$
	# Tokens	Max $256\times(1+9)$	Max $256\times(1+9)$	Max $256\times(1+9)$
Data	**Dataset**	LLaVA-Pretrain	Enhancement (Tab. 1)	SFT (Tab. 2)
	# Samples	558K	1.8M	3M
Training	**Trainable**	Projector	ViT & Projector & LLM	SRA & Projector & LLM
	Load SRA	✗	✗	✓
	Batch Size	256	256	256
	LR: LLM	2×10^{-5}	2×10^{-5}	2×10^{-5}
	LR: Projector	1×10^{-3}	2×10^{-5}	2×10^{-5}
	LR: ViT / SRA	-	2×10^{-6}	2×10^{-4}
	Epoch	1	1	1

Table 3. **Detailed configuration for three-stage training of HiRes-LLaVA.** The table illustrates the vision configurations, dataset characteristics, and training hyperparameters.

Model	Data	VQA-Text	ChartQA	DocQA	InfoVQA	MMB	MME-P
LLaVA-1.5	LLaVA-665k	53.3	13.7	14.2	19.4	71.1	1459.66
LLaVA-1.5	LLaVA-665k + Doc-79k	53.3	23.8	22.6	31.4	70.7	1424.6
HiRes-LLaVA	LLaVA-665k	62.4	19.8	37.7	26.0	72.3	1486.1
HiRes-LLaVA	LLaVA-665k + Doc-79k	62.3	57.6	58.5	39.2	71.1	1444.8

Table 4. Ablation study of different training data. Using the same training data, our HiRes-LLaVA consistently outperforms LLaVA-1.5, demonstrating the superior effectiveness of our approach.

Type	VQA-Text	ChartQA	DocQA	InfoVQA	MMB	MME-P
Same	57.2	39.7	52.6	37.6	61.3	1379.8
Separated	**61.8**	**58.8**	**59.7**	**41.4**	**65.5**	**1456.1**

Table 5. Ablation of the separator. 'Separated' means three separators are the difference and 'Same' means that three separators are the same.

LLaVA-1.5 under these two training data sets, confirms that the superior performance can be attributed to the method itself rather than the volume of data.

Ablation of the separators To further evaluate the effect of the separators, we conduct experiments on whether the separators are different or the same. Tab. 5 demonstrates that using separated separators greatly outperforms using the same ones which would confuse the model about the position of slices.

C. Efficiency Analysis

Comparison with other LVLMs. To validate the efficiency of our method, we compare the computational cost, training, and inference times with various LVLMs in Appendix C. For computational cost, we report the FLOPs of the ViT backbone, connector, and LLM components for each model. Experimental results demonstrate that HiRes-LLaVA, despite processing inputs at twice the resolution of LLavA-Next (1344^2 vs. 672^2), is able to reduce training time by approximately 74%.

Comparison with other downsampling methods. We also compare the FLOPs and training time of our proposed downsampling strategy SMS with other vision token downsamplers, including ConcatChannel [6], Q-Former [2], and C-Abstractor [4], as shown in Tab. 7. The results show that our SMS, even when combined with additional components like SRA, achieves competitive efficiency compared to existing state-of-the-art downsamplers.

D. Discussion

What's the goal of the EntityGrid-QA benchmark? The goal of our EntityGrid-QA benchmark is to assess the fragmentation issue in LVLMs (Large Vision-Language Models) when processing high-resolution inputs, rather than their ability to identify different types of objects. To address this, EntityGrid-QA synthesizes images by iteratively placing objects in different positions, allowing us to evaluate how these

Training batch size	Inference Resolution	*FLOPs* ViT	Connector	LLM	Training time	Inference time
HiRes-LLaVA						
2	1344x1344	6.6 T	195.2 G	37.1 T	60.7h **(15.9%)**	15.4m
HiRes-LLaVA w/o SRA						
2	1344x1344	6.5 T	195.2 G	37.1 T	59.5h **(15.6%)**	12.9m
LLaVA-Next (LLaVA-1.6)						
2	1344x1344	Out of the memeory				
1	672x672	1.9 T	120.8 G	44.0 T	381.0h	13.2m

Table 6. Comparison of the efficiency of different models. Note that training time is assessed under the SFT setting on a machine with 8 V100 GPUs. The inference time is assessed on the InfoVQA benchmark with 6096 images by using the lmms-eval. Note that using the same batch size per device and resolution, LLaVA-Next would be out of the memory. The ratios of training time for ours relative to LLaVA-Next are marked in **purple**.

Components Downsampler	SRA	*FLOPs* ViT	Sampler	LLM	Training Time
NoDownsample	✗	6.5 T	410.8 G	148.3T	-
ConcatChannel	✗	6.5 T	164.3 G	37.1 T	58.6h
Q-Former	✗	6.5 T	205.5 G	37.1 T	58.9h
C-Abstractor	✗	6.5 T	258.2 G	37.1 T	60.7h
SMS	✗	6.5 T	195.2 G	37.1 T	59.5h
SMS	✓	6.6 T	195.2 G	37.1 T	60.7h

Table 7. Ablation study of the efficiency of individual components for different downsamplers. We assume the inputs are an image with 16 slices and 100 text tokens. Note that no downsampling method causes out-of-memory (OOM) issues during training. Training time is assessed under the SFT setting on a machine with 8 V100 GPUs.

Benchmarks	Slicing Strategy	Target Issue
LLaVA-UHD's	Overlapped	Counting
Our EntityGrid-QA	Non-overlapped	Fragmentation

Table 8. The differences between our EntityGrid-QA and LLaVA-UHD's benchmark [11].

models perform on the edges and the center of the slices. Compared to harvesting real-world images with answer targets on the edges of slices, the synthesized approach is more simple-to-collect, effective, flexible, sufficient to evaluate the fragmentation issue.

Compared with LLaVA-UHD. The target issues and slicing strategies are different between Hires-LLaVA and LLaVA-UHD [11]. While LLaVA-UHD reveals the counting problem in the overlap slicing strategy for the high-resolution image inputs, Hires-LLaVA focuses on the fragmentation issues of non-overlapped slicing strategy which is commonly used in recent open-sourced high-resolution LVLMs. Ta-

ble 8 summarize the differences of our EntityGrid-QA and LLaVA-UHD's benchmark.

E. More Visualization

Samples from EntityGrid-QA Benchmark. We illustrate three examples from our proposed EntityGrid-QA benchmark in Fig. 1. These four samples visualize examples of the four tasks in the benchmark we proposed. For each task, we write or paste the digital number or object directly onto each position of an empty image, and ask questions to the models.

More Qualitative Results. To further validate the effectiveness of our model, we illustrate the more qualitative results of InfoVQA, ChartQA and V* Benchmark in Fig. 2 and Fig. 3. Moreover, we give two qualitative examples to present the HiRes-LLaVA's capability of generating HTML code when given a website image in Fig. 4.

F. Broader Impacts

The development of HiRes-LLaVA advances the field of vision-language models and has broad implications for various applications, including document analysis, medical imaging and remote sensing. However, alongside these potential benefits, there are considerable concerns.

HiRes-LLaVA, not having undergone rigorous safety training, might generate harmful or inappropriate content, leading to legal and ethical issues. Furthermore, its enhanced ability to process high-resolution inputs could be misused for creating misleading news, contributing to disinformation. These potential negative impacts highlight the need for careful management and ethical guidelines in the deployment of such technologies.

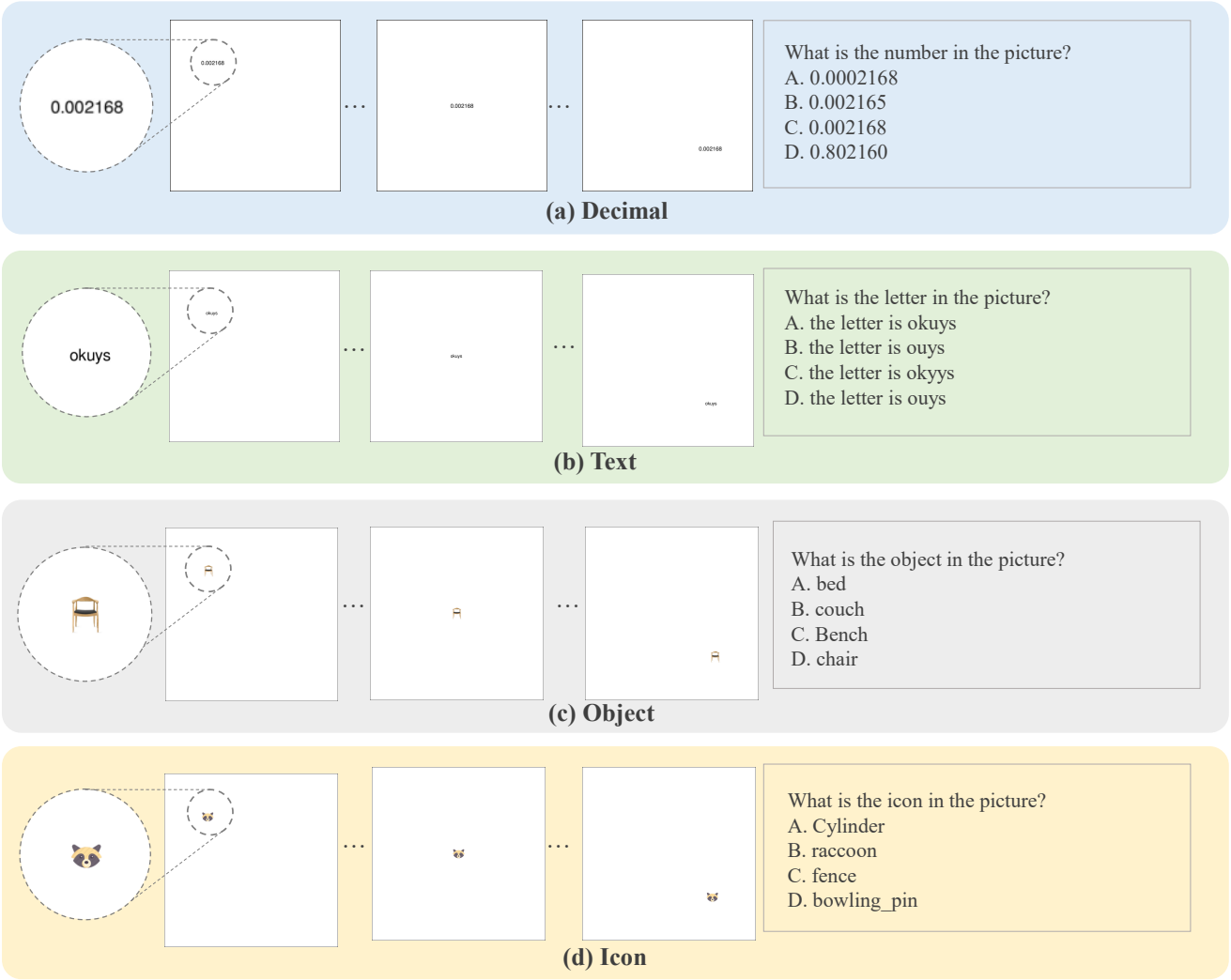
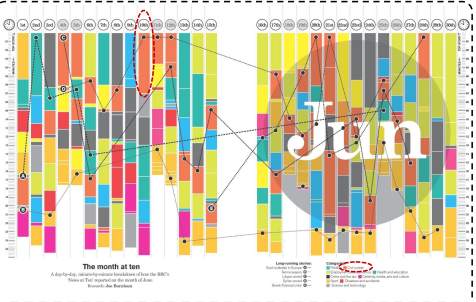


Figure 1. **Examples of our proposed EntityGrid-QA Benchmark.**

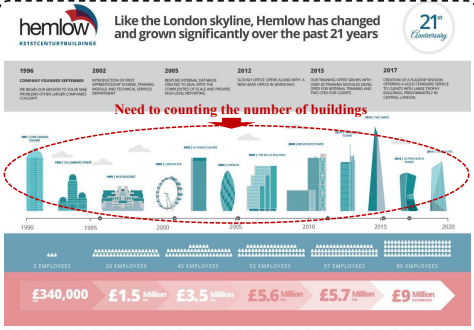


Q: What category was the ***top story*** on the ***10th*** of June?

LLaVA-Next: politics ✗ **Monkey**: politics ✗
Ours: civil unrest √

Q: Which environment issue is mentioned ***in the bottom row of the bulb*** image in the infographic?

LLaVA-Next: waste ✗ **Monkey**: pollution ✗
Ours: ocean pollution √

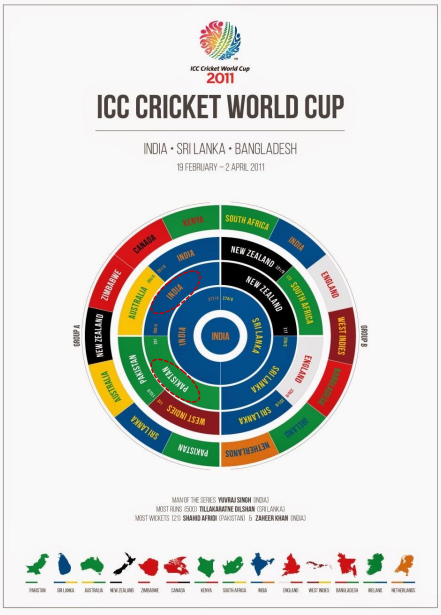


Q: ***How many*** buildings were constructed by Hemlow?

LLaVA-Next: 21 ✗ **Monkey**: 9 ✗
Ours: 12 √

Q: What ***is the third ingredient*** listed to make Pasta?

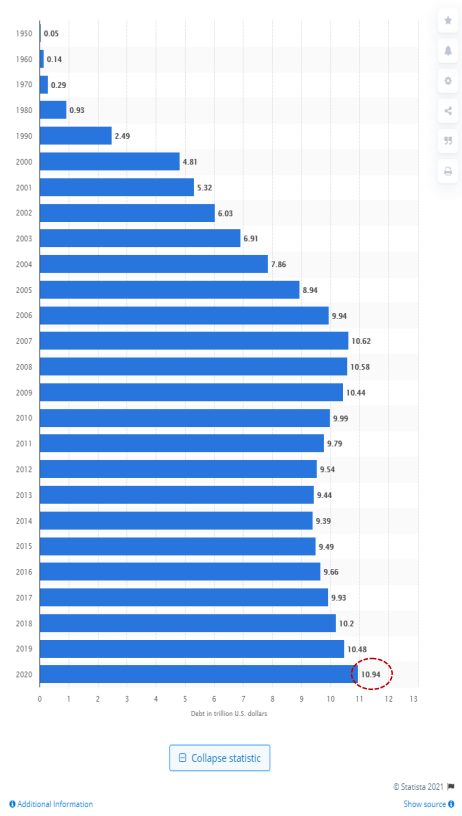
LLaVA-Next: water ✗ **Monkey:** eggs ✗
Ours: salt √



Q: Who was the ***opponent of India*** in the ***semifinals*** of World Cup 2011?

LLaVA-Next: england ✗ **Monkey**: sri lanka ✗
Ours: pakistan √

Figure 2. **Qualitative results from InfoVQA [31].**

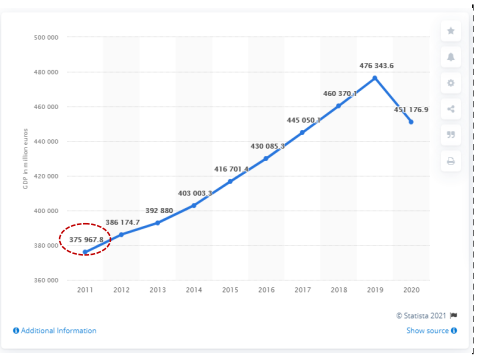


Q: What was the mortgage debt in the United States in 2020?

LLaVA-Next: 10.4 ✗ **Monkey**: 12.13 ✗
Ours: 10.94 √

Q: What is the color of ***the woman's scarf***?
(A) white
(B) red
(C) yellow
(D) green

LLaVA-Next: A ✗ **Monkey**: A ✗
Ours: B √



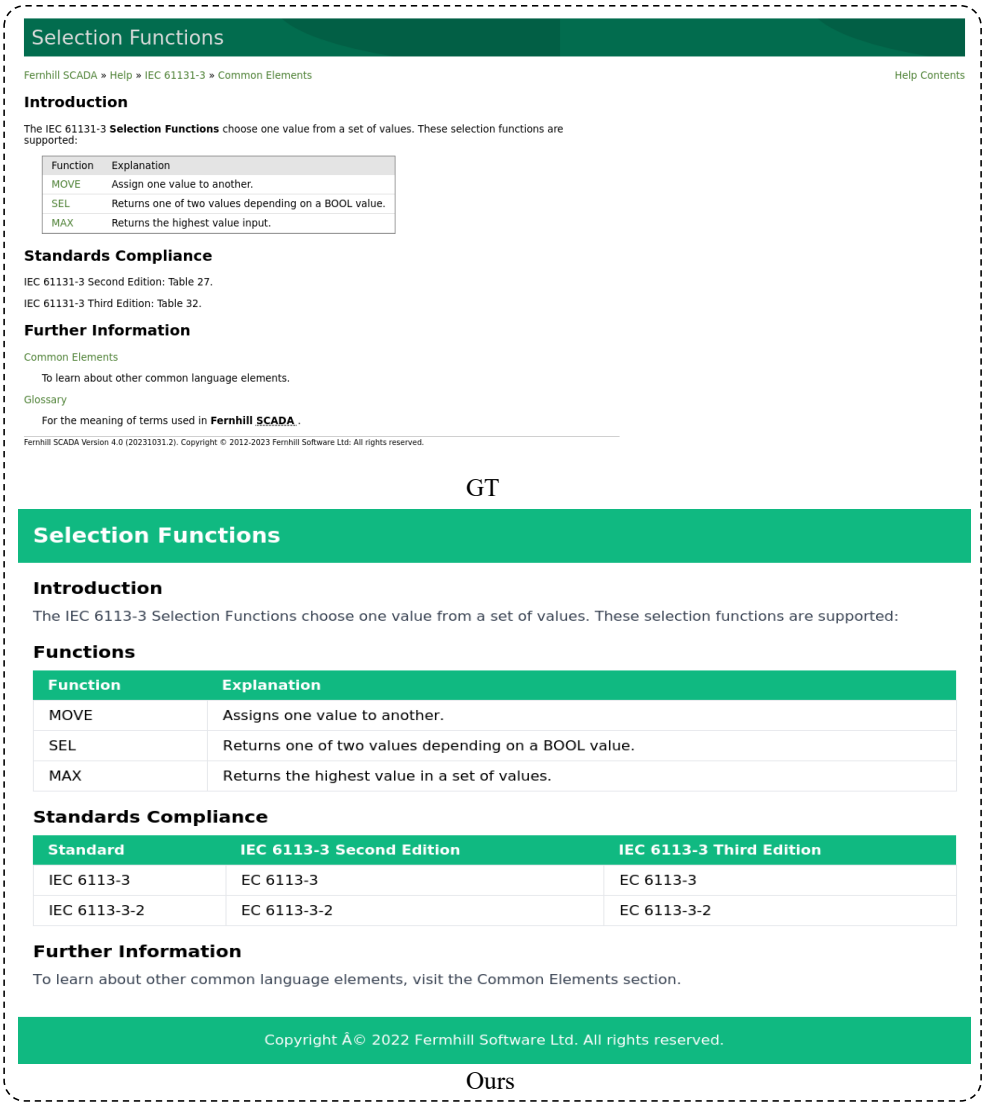
Q: What was Belgium's GDP in 2011?

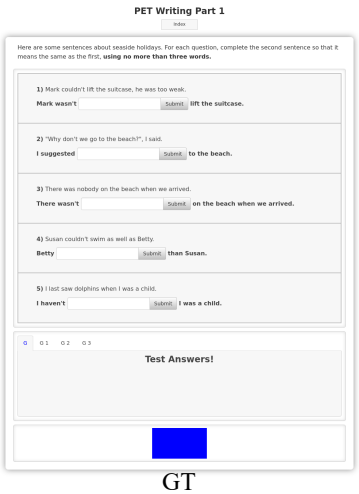
LLaVA-Next: 37596.8 ✗
Monkey: 386174.7 ✗
Ours: 375967.8 √

Q: What is the ***cartoon character*** on the clock?
(A) Bugs Bunny
(B) Mickey Mouse
(C) SpongeBob
(D) Donald Duck

LLaVA-Next: C ✗ **Monkey**: A ✗
Ours: B √

Figure 3. **Qualitative results from ChartQA [30] and Vstar Benchmark [39].** We use the red circle to highlight the answer target in the image.





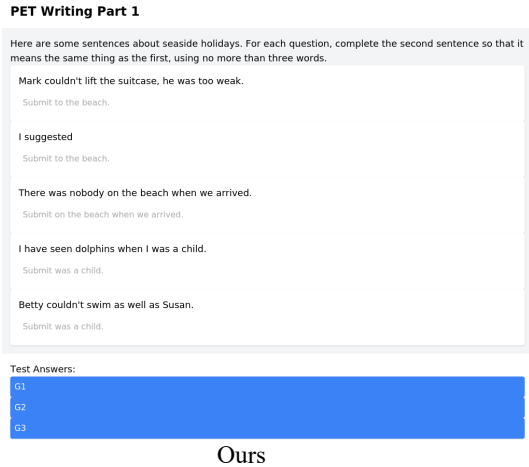


Figure 4. **Qualitative results on Image2HTML task [34].** We visualize convert the generated html code to website image and compare to the input image.

References

[1] 100TAL. TAL Education Group. https://ai.100tal.com/dataset, 2023. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 3

[3] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4291–4301, 2019. 1

[4] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[5] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model, 2024. 1

[6] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv:2310.09478*, 2023. 3

[7] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024. 1

[8] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019. 1

[9] Xingyu Chen, Zihan Zhao, Lu Chen, Jiabao Ji, Danyang Zhang, Ao Luo, Yuxuan Xiong, and Kai Yu. Websrc: A dataset for web-based structural reading comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4173–4185, 2021. 1

[10] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In *ACL*, pages 845–855, 2018. 1

[11] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer, 2024. 3

[12] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, pages 5648–5656, 2018. 1

[13] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pages 235–251, 2016. 1

[14] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *CVPR*, pages 4999–5007, 2017. 1

[15] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *ECCV*, 2022. 1

[16] Shanghai AI Laboratory. Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o, 2023. 1

[17] LAION. Gpt-4v dataset. https://huggingface.co/datasets/laion/gpt4v-dataset, 2023. 1

[18] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. 1

[19] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *SIGIR*, pages 3108–3120, 2022. 1

[20] Bo Li*, Peiyuan Zhang*, Kaichen Zhang*, Fanyi Pu*, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimoal models, 2024. 1

[21] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. 1

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 1

[23] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *TACL*, 11:635–651, 2023. 1

[24] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 1

[25] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*, 2023. 1

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2023. 1

[27] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1

[28] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021. 1

[29] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 35: 2507–2521, 2022. 1

[30] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, pages 2263–2279, 2022. 1, 7

[31] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022. 1, 6

[32] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *WACV*, pages 1527–1536, 2020. 1

[33] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pages 947–952, 2019. 1

[34] Chenglei Si, Yanzhe Zhang, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2code: How far are we from automating front-end engineering?, 2024. 8

[35] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 1

[36] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 1

[37] S Svetlichnaya. Deepform: Understand structured documents at scale, 2020. 1

[38] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024. 1

[39] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024. 7

[40] Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*, 2024. 1

[41] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023. 1

[42] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866, 2022. 1