LocateAnything3D: Vision-Language 3D Detection with Chain-of-Sight

Supplementary Material

Supplementary Contents

A. Additional Experiments and Analysis

A.1. Quantitative Evaluation of 3D Grounding

Problem setting. To further evaluate our LocateAnything3D's capability in following spatial language instructions, we conduct experiments on indoor 3D grounding benchmarks. We strictly follow the experimental protocol established by Cube-LLM [21]. Specifically, we repurpose the test sets of three standard indoor detection datasets: *Objectron* [2], *ARKitScenes* [8], and *SUN-RGBD* [114] into grounding benchmarks. The task requires the model to localize particular objects based on text prompts that vary in specificity: (1) Category-only: The prompt contains only the object class name (*e.g.*, "chair"); and (2) Category+Location: The prompt includes the class name augmented with spatial descriptions derived from the object's position relative to the camera (*e.g.*, "chair on the left", "bookshelf close to camera"). The spatial qualifiers (left/right/center and close/medium/far) are generated based on the 2D image coordinates and depth thresholds defined in the baseline setting. We report the Average Precision ($\mathrm{AP}_{\mathrm{3D}}$) averaged over $\mathrm{IoU}_{\mathrm{3D}}$ thresholds of $\tau \in \{0.15, 0.25, 0.50\}$. If multiple objects match the text description, the maximum IoU among them is used for evaluation.

Evaluation results. Table A summarizes the results on the benchmark. We copy the Cube-LLM numbers for models pre-trained on the "LV3D-small" and full "LV3D" corpora from their paper [21], and add our LocateAnything3D model, which is trained using the Chain-of-Sight formulation on our unified 3D corpus. Across all three datasets and both metrics, LocateAnything3D substantially outperforms Cube-LLM, despite no task-specific architecture changes for indoor scenes.

From the table, we can also notice that Cube-LLM [21] achieves lower performance for $\mathrm{AP}_{\mathrm{3D}}^{\mathrm{cat+loc}}$ than $\mathrm{AP}_{\mathrm{3D}}^{\mathrm{cat}}$, in two out of the three evaluation scenarios. On the contrary, LocateAnything3D achieves consistent performance improvement when location information is provided to the model as additional conditions. This different clearly highlights the higher capability of our model to interpret spatial descriptions and 3D understanding.

Problem with point-cloud grounding benchmarks. Existing indoor 3D grounding datasets such as ScanRefer [16] and ReferIt3D [1] are explicitly built around point clouds rather than images, and are therefore ill-suited to our monocular 3D detection setting. Each scene in these benchmarks is represented by a single reconstructed point cloud but is associated with many RGB views that only partially observe the scene. Referring expressions are written to identify objects in the global 3D scene, not in a particular camera view, and a single object may be visible in multiple images with very different appearances and levels of occlusion. As a result, there is no unambiguous way to assign a unique image and 3D box pair to each language query, and any attempt to project the point-cloud annotations into 2D would depend on arbitrary choices of viewpoint and visibility thresholds. For this reason, we follow Cube-LLM [21] and evaluate our model on indoor benchmarks derived from Objectron [2], ARKitScenes [8], and SUN-RGBD [114], where each image already comes with camera-specific 3D boxes and thus naturally supports monocular 3D detection and grounding.

A.2. Data Efficiency and Training Dynamics

To better understand the contributions of our Chain-of-Sight (CoS) formulation and the role of 2D pretraining, we conduct a detailed analysis of our model's performance under limited data regimes and different initialization strategies. Figure A visualizes these comparisons.

Impact of chain-of-sight on data efficiency. Figure A (left) compares our full 2D-3D CoS formulation against a "pure 3D" decoder trained without any explicit 2D step. On the horizontal axis we vary the fraction of our 3D training corpus from 10% to 100%; the dashed line marks the performance of DetAny3D [153] (32.2 $\mathrm{AP}_{\mathrm{3D}}$).

Across all data regimes, the CoS model is consistently

Table A. **Indoor 3D Object Grounding Performance.** We compare LocateAnything3D against Cube-LLM trained on different data scales. Cube-LLM$_{\text{small}}$ is trained on the LV3D-small subset, while Cube-LLM$_{\text{large}}$ is trained on the full LV3D dataset containing approximately **9.6M images**. In contrast, our model is trained on a much smaller curated dataset of **1.7M images**. Despite this significant disparity in data scale, LocateAnything3D outperforms the best baseline by a large margin across all benchmarks. We report Average Precision (AP) prompted with either category names ($\text{AP}_{\text{3D}}^{\text{cat}}$) or category plus spatial location ($\text{AP}_{\text{3D}}^{\text{cat+loc}}$).

Method	Objectron		ARKitScenes		SUN-RGBD	
	$\text{AP}_{\text{3D}}^{\text{cat}}$	$\text{AP}_{\text{3D}}^{\text{cat+loc}}$	$\text{AP}_{\text{3D}}^{\text{cat}}$	$\text{AP}_{\text{3D}}^{\text{cat+loc}}$	$\text{AP}_{\text{3D}}^{\text{cat}}$	$\text{AP}_{\text{3D}}^{\text{cat+loc}}$
Cube-LLM$_{\text{small}}$ [21]	56.7	36.1	21.6	28.3	25.5	25.5
Cube-LLM$_{\text{large}}$ [21]	69.8	45.4	23.5	31.8	29.7	28.8
LocateAnything3D (Ours)	**72.5**	**75.0**	**41.7**	**53.9**	**29.7**	**39.5**
Δ vs. Cube-LLM$_{\text{large}}$	+2.7	+29.6	+18.2	+22.1	+0	+10.7

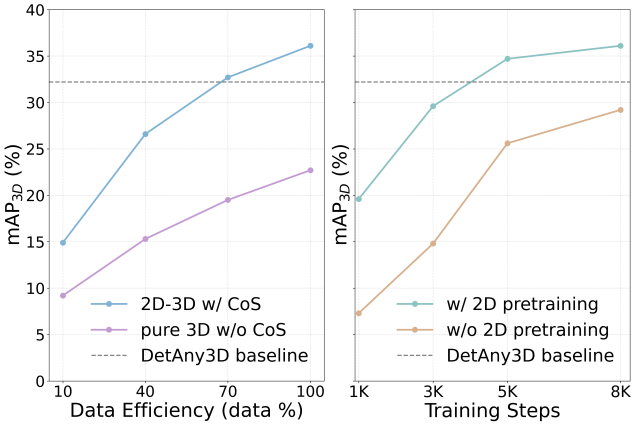


Figure A. **Data efficiency and training dynamics analysis**. (1) The left figure shows data efficiency: We report AP_{3D} vs. percentage of training data used. Our Chain-of-Sight (CoS) formulation (blue) consistently outperforms direct 3D prediction (purple), achieving competitive performance with only 10% of the data. (2) The right figure shows training dynamics: We compare training curves with and without 2D detection pretraining. 2D pretraining (green) accelerates convergence significantly, surpassing the previous state of the art (dashed line) almost immediately, whereas training from scratch (orange) is slower and yields lower final accuracy.

stronger and markedly more data-efficient than the pure 3D variant. With only 10% of the data, CoS outperforms pure 3D prediction baseline by a large margin. As we further scale the data to 70% and 100%, the CoS curve continues to climb to 32.7 and 36.1 AP_{3D}, whereas pure 3D saturates at 19.5 and 22.7. This supports our central claim that explicitly factorizing 3D detection into a 2D grounding step followed by 3D lifting is not just more accurate, but also significantly more sample-efficient.

Impact of 2D pretraining on convergence. Figure A (right) studies the effect of the 2D grounding pretraining stage on our training dynamics. We plot AP_{3D} as a function of CoS training steps, comparing models initialized with and without 2D pretraining, and again mark the DetAny3D performance with a dashed line.

Initializing from the 2D grounding stage yields a substantial head start. After only 1k CoS steps, the pretrained model already achieves 19.6 AP_{3D}, whereas the model trained from scratch is still at 7.3. As training proceeds, both curves improve, but the gap persists. At the final checkpoint, the model with 2D pretraining converges to 36.1 AP_{3D}, while the scratch model lags behind at 29.2. This indicates that robust 2D localization capabilities serve as a critical foundation for 3D perception, allowing the model to focus its capacity on lifting 2D features to 3D space rather than learning basic localization from scratch.

A.3. Impact of Token Serialization Strategy

To further validate our Chain-of-Sight (CoS) design of interleaving 2D and 3D tokens on the per-object level ($\text{2D}_{\text{i}} \rightarrow \text{3D}_{\text{i}}$), we compare it against a "clustered" decoding strategy. In the clustered setting, the model is trained to predict all 2D bounding boxes for the scene first, followed by all corresponding 3D bounding boxes ($\text{2D}_{\text{1...N}} \rightarrow \text{3D}_{\text{1...N}}$). This ablation tests whether the tight coupling of 2D visual evidence with its corresponding 3D geometry is necessary, or if the model can simply learn two separate phases of detection. We report results trained for 1 epoch on three distinct datasets to analyze performance across different scene complexities. As shown in Table B, our interleaved default setting consistently outperforms the clustered strategy. The magnitude of this performance gap is strongly correlated with scene clutter and object density.

Object-centric scenes. On Objectron, which typically contains only 1 or 2 prominent objects per image, the performance gap is minimal (61.5 vs. 63.0). The additional effort for the model to associate the i-th 3D box with the i-th 2D box is negligible.

Structured outdoor scenes. KITTI scenes contain more objects with large depth range, but they follow a structured distribution (cars on a road) with relatively clear depth ordering. While the gap widens, the model can still maintain reasonable 2D-3D association in the clustered setting.

Table B. **Ablation of Token Serialization Strategy.** We compare our default *Interleaved* Chain-of-Sight strategy ($2D_i \rightarrow 3D_i$) against a *Clustered* strategy where all 2D boxes are predicted before all 3D boxes ($\mathrm{2D_{1...N}} \rightarrow \mathrm{3D_{1...N}}$) with average precision ($\mathrm{AP_{3D}}$). Models are trained for 1 epoch. The results show that the interleaved strategy is significantly more robust, especially in cluttered scenes where associating separated 2D and 3D sequences becomes difficult.

Serialization Strategy	Average Precision ($\mathrm{AP_{3D}}$)		
	Objectron	KITTI	Hypersim
Clustered ($\mathrm{2D_{1...N}} \rightarrow \mathrm{3D_{1...N}}$)	61.5	17.4	4.7
Interleaved ($\mathrm{2D_i} \rightarrow \mathrm{3D_i}$, **Ours**)	**63.0**	**22.1**	**11.2**
Performance Gap	+1.5	+4.7	+6.5

Highly cluttered scenes. The most significant drop occurs on Hypersim which is characterized by chaotic indoor scenes with dozens of objects and frequent occlusions. In these scenarios, the clustered strategy fails catastrophically. The model struggles to maintain the implicit alignment between the k-th 2D box generated early in the sequence and the k-th 3D box generated much later, resulting in a big difference between the two setting.

A.4. Runtime Analysis

Although LocateAnything3D is primarily designed as a general 3D perception VLM rather than a real-time perception system, we report its end-to-end inference latency for completeness. On average, processing a single image-query pair with LocateAnything3D takes **683 ms** under our evaluation setup with a single H100 GPU. This wall-clock time consists of three main components: (1) vision encoding of the input image, (2) LLM pre-filling with the textual prompt, and (3) autoregressive generation of the mixed 2D/3D box tokens produced by the Chain-of-Sight decoder.

To isolate the cost of the Chain-of-Sight factorization, we compare our full 2D-3D CoS model with a pure-3D variant that directly predicts 3D boxes without emitting intermediate 2D boxes. Introducing the 2D step increases the average latency by only **121 ms** (from roughly 562 ms to 683 ms), yet enables the substantial accuracy gains and data-efficiency improvements, as reported in the Section 5 of the main paper. In other words, CoS adds a modest computational overhead while making 3D detection both easier to learn and significantly more accurate.

We emphasize that LocateAnything3D is not meant to replace highly optimized real-time detectors used in latency-critical loops (*e.g.*, onboard obstacle avoidance). Instead, our goal is to endow a general-purpose VLM with strong 3D grounding capabilities so that it can serve as a foundation for downstream tasks such as offline planning, scene understanding, and multimodal agent reasoning. In this context, a sub-second per-image latency is well within an acceptable range, especially given the unified interface and performance benefits brought by the Chain-of-Sight formulation.

B. Implementation Details

B.1. Models, Tokenization, and Prompting

Model designs. (1) vision encoder. We use SigLIP [150] with FlashAttention 2 [24] enabled. (2) Language model. We use a Qwen2 8B causal LM [122] with FlashAttention 2, trained end-to-end (no freezing). (3) Multimodal connector. We use an MLP projector, which maps SigLIP tokens to the LLM hidden space with two-layer MLP.

Image tokenization. A tiling-based tokenization where we decompose images into patches of a forced image size of 448. The total image tokens scale linearly with the number of tiles.

Conversation format. Qwen2-chat template. Image tokens are inserted by replacing each `<image>` placeholder with `<IMG_START>` followed by repeated `<IMG_CONTEXT>` tokens and `<IMG_END>`. The repeat count equals per-tile tokens times the number of tiles for that image; we assert a strict match between precomputed and actual counts.

Labels. Only assistant spans are supervised; all instruction tokens are masked. Truncation safety checks keep training targets valid.

B.2. Dynamic Tiling and Packing

Tiling and image processing. Images are decomposed into an adaptive grid of 448-pixel tiles, min 1 and max 12 tiles, plus an optional global thumbnail. Tiling policy selects the closest aspect ratio while favoring large area coverage for stability.

Sequence construction and online packing. Our context length is 16,384 tokens per sample. We enable online packing to concatenate multiple short samples until the context budget is filled while tracking sub-sample boundaries in the attention mask. A dummy image is inserted only if the entire packed sample is text-only. Position ids respect packed boundaries; the model supports sequence parallel groups but we run with degree 1 in our experiments.

B.3. Optimization and Systems

We use a precision of bfloat16 across vision and language. For memory handling, gradient checkpointing is enabled for both the SigLIP encoder and the LLM; fused ops are used to reduce memory overhead. For the loss function, we fuse the linear cross-entropy with per-sample normalization using the number of valid answer tokens. For the optimizer and schedule, we use AdamW with a learning rate of 1e-5, a weight decay of 0.05, a cosine decay, and a warm-up of 3%.

Category	Dataset
Captioning & Knowledge	ShareGPT4o [98], KVQA [108], Movie-Posters [113], Google-Landmark [132], WikiArt [40], Weather-QA [81], Coco-Colors [34], music-sheet [27], SPARK [146], Image-Textualization [100], SAM-Caption [101], Tmdb-Celeb-10k [4]
Mathematics	GeoQA+ [13], MathQA [143], CLEVR-Math/Super [63, 67], Geometry3K [75], MAVIS-math-rule-geo [155], MAVIS-math-metagen [155], InterGPS [76], Raven [152], GEOS [107], UniGeo [18]
Science	AI2D [49], ScienceQA [78], TQA [50], PathVQA [36], SciQA [5], Textbooks-QA, VQA-RAD [57], VisualWebInstruct [124]
Chart & Table	ChartQA [85], MMC-Inst [69], DVQA [45], PlotQA [89], LRV-Instruction [68], TabMWP [79], UniChart [86], Vistext [120], TAT-DQA [163], VQAonBD [128], FigureQA [46], Chart2Text [48], RobuT-{Wikisql, SQA, WTQ} [159], MultiHiertt [158]
Naive OCR	SynthDoG [53], MTWI [35], LVST [117], SROIE [39], FUNSD [43], Latex-Formula [97], IAM [84], Handwriting-Latex [3], ArT [20], CTW [148], ReCTs [154], COCO-Text [127], SVRD [145], Hiertext [73], RoadText [125], MapText [64], CAPTCHA [99], Est-VQA [130], HME-100K [118], TAL-OCR-ENG [118], TAL-HW-MATH [118], IMGUR5K [54], ORAND-CAR [25], Invoices-and-Receipts-OCR [95], Chrome-Writting [93], IIIT5k [90], K12-Printing [118], Memotion [104], Arxiv2Markdown, Handwritten-Mathematical-Expression [6], WordArt [134], RenderedText [131], Handwriting-Forms [42]
OCR QA	DocVQA [22], InfoVQA [88], TextVQA [112], ArxivQA [62], ScreencQA [38], DocReason [94], Ureader [140], FinanceQA [116], DocMatrix [58], A-OKVQA [106], Diagram-Image-To-Text [47], MapQA [15], OCRVQA [91], ST-VQA [10], SlideVQA [119], PDF-VQA [26], SQuAD-VQA, VQA-CD [82], Block-Diagram [110], MTVQA [121], ColPali [29], BenthamQA [87]
General VQA	LLaVA-150K [70], LVIS-Instruct4V [129], ALLaVA [17], Laion-GPT4V [56], LLAVAR [157], SketchyVQA [126], VizWiz [33], IDK [14], AlfworldGPT, LNQA [103], Face-Emotion [28], SpatialSense [138], Indoor-QA [51], Places365 [161], MMinstruct [71], DriveLM [111], YesBut [96], WildVision [80], LLaVA-Critic-113k [135], RLAIF-V [144], VQAv2 [31], MMRA [133], KONIQ [37], MMDU [72], Spot-The-Diff [44], Hateful-Memes [52], COCO-QA [105], NLVR [115], Mimic-CGD [59], Datikz [9], Chinese-Meme [23], IconQA [77], Websight [60]
Text-only	Orca [65], Orca-Math [92], OpenCodeInterpreter [160] MathInstruct [149], WizardLM [136], TheoremQA [19], OpenHermes2.5 [123], NuminaMath-CoT [61], Python-Code-25k [30], Infinity-Instruct [7], Python-Code-Instructions-18k-Alpaca [41], Ruozhiba [74], Infinity-MATH [151], StepDPO [55], TableLLM [156], UltraInteract-sft [147]
2D Grounding & Counting	RefCOCO/+/g (en) [83, 142], Objects365 [109], COCO [66], EgoObjects [162], BLIP3-OCR [137], BDD100K [141], Nuimages [11], Flick30K [102], LVIS [32]

Table C. **Summary of our extensive and diverse supervised fine-tuning datasets for 2D pretraining.** We use a comprehensive collection of numerous large-scale datasets spanning multiple domains and tasks to pretrain our model, ensuring broad coverage and robust performance across diverse visual and language understanding scenarios.

Packing target. We use dynamic online packing to saturate the 16K context; the scripts set an iteration-level token target of 2^{17} ($= 128K$) to govern accumulation and throughput.

Training scale. We train our model using 64 H100 GPUs. The whole training takes 46 hours with 37K steps, distributed with torchrun and DeepSpeed ZeRO-3.

B.4. 2D Grounding Pretraining

Dataset composition. We pretrain on a large-scale 2D grounding corpus covering four domains with different data mixture percentage: (1) **General detection:** Object365 [109] (5 epochs), COCO [66] (12 epochs), and LVIS [32] (3 epochs); (2) **Ego-centric & driving:** BDD100K [141] (3 epochs), nuImages [12] (3 epochs), and EgoObjects [162] (3 epochs); (3) **Referring-expression grounding:** RefCOCO [142] (3 epochs), RefCOCO+ [83] (3 epochs), RefCOCOg [83] (3 epochs), and Flickr30k [102] (3 epochs); (4) **Text grounding:** a BLIP3-OCR subset [137] ($\approx$ 1.0M samples). Overall, this results in over **15M** multi-turn dialogues in the grounding corpus, which we mix with an additional **8M** samples for general instruction tuning, as demonstrated in Table C.

Annotation format. For each image, we construct a multi-turn dialogue where each turn follows the instruction template "Detect all the objects in the image that belong to the category set {c}." The response is either a comma-separated list of 2D bounding boxes in $[x_1, y_1, x_2, y_2]$ format (top-left to bottom-right, integer-quantized to $[0, 1000]$), or "None" if no instance exists. We include all positive categories present in the image and sample 10 absent categories as negatives, yielding per-image supervision that mixes existence and non-existence signals across multiple dialogue turns.

C. Limitations and Future Work

While LocateAnything3D establishes a strong foundation for VLM-native 3D perception, several avenues remain for future exploration. Our work primarily focuses on validating the Chain-of-Sight (CoS) decoding mechanism within a single-frame, end-to-end setting. Below, we outline key directions where our framework can be naturally extended to incorporate additional geometric signals and temporal contexts.

Integration of explicit depth priors. Currently, our model infers metric depth solely from monocular RGB cues and semantic context. While the near-to-far curriculum effectively regularizes this process, the model does not yet leverage explicit depth maps. Future work could introduce a depth encoder or use depth images as an additional conditioning input. This would allow the model to utilize output from state-of-the-art monocular depth estimators as a geometric prompt, potentially improving metric accuracy in texture-less or ambiguous scenes.

Explicit camera intrinsic conditioning. Our current approach normalizes 3D coordinates into a unified camera-centric space to maximize cross-dataset generalization. However, it implicitly relies on the vision encoder to handle variations in focal length and field of view. An extension is to explicitly tokenize camera intrinsic matrices (*e.g.*, focal length, principal point) and feed them as positional prompts. This would allow the decoder to mathematically adjust its size and depth predictions based on the specific camera optics, rather than learning an average projection model.

Extension to multi-frame and video settings. The cur-

rent framework operates on single images. However, the autoregressive nature of our decoder is naturally suited for temporal sequences. Future iterations could extend the context window to include visual tokens from preceding frames. The model could learn to track objects over time, estimate velocity, and leverage multi-view consistency to resolve depth ambiguities that are ambiguous in a single frame.

D. Broader Impact

The development of LocateAnything3D represents a step toward unifying semantic understanding and metric perception within general-purpose foundation models. By enabling VLMs to perceive the physical world in 3D without specialized heads, we lower the barrier to entry for developing capable embodied agents and home robotics. This has positive implications for industries ranging from autonomous driving to assistive robotics.

However, we acknowledge potential risks associated with this technology. Like all deep learning models trained on web-scale data, our model may inherit biases present in the training corpora, such as geographic or cultural biases in object distributions. This could lead to uneven performance across different regions. We encourage the research community to prioritize the development of diverse, representative 3D datasets and to consider the ethical implications of spatial intelligence in deployment scenarios.

E. More Case Visualization

We provided more qualitative visualization in this section.

Failure case visualization. Figure B demonstrates some representative failure cases of our method. Despite the great performance, our model still suffers from the lack of diverse and high-quality 3D annotations compared to the 2D scenario. Hence, similar to the baselines [139, 153], it faces challenges when presented with scenes that exhibit very different focal length, spatial layouts, and textural details.

More successful visualization. Figure C demonstrates more successful cases.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, 2020. 1

[2] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *CVPR*, 2021. 1

[3] aidapearson. Aida calculus math handwriting recognition dataset. https://www.kaggle.com/datasets/aidapearson/ocr-data, 2023. 4

[4] Ashraq. Tmdb-celeb-10k dataset. https://huggingface.co/datasets/ashraq/tmdb-celeb-10k, 2024. 4

[5] Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13 (1):7240, 2023. 4

[6] Azu. Handwritten-mathematical-expression-convert-latex. https://huggingface.co/datasets/Azu/Handwritten-Mathematical-Expression-Convert-LaTeX, 2023. 4

[7] BAAI. Infinity-instruct dataset. https://huggingface.co/datasets/BAAI/Infinity-Instruct, 2024. 4

[8] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 1

[9] Jonas Belouadi, Anne Lauscher, and Steffen Eger. Automatikz: Text-guided synthesis of scientific vector graphics with tikz. *arXiv preprint arXiv:2310.00367*, 2023. 4

[10] 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. 4

[11] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 4

[12] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 4

[13] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *COLING*, pages 1511–1520, 2022. 4

[14] Sungguk Cha, Jusung Lee, Younghyun Lee, and Cheoljong Yang. Visually dehallucinative instruction generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5510–5514. IEEE, 2024. 4

[15] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022. 4

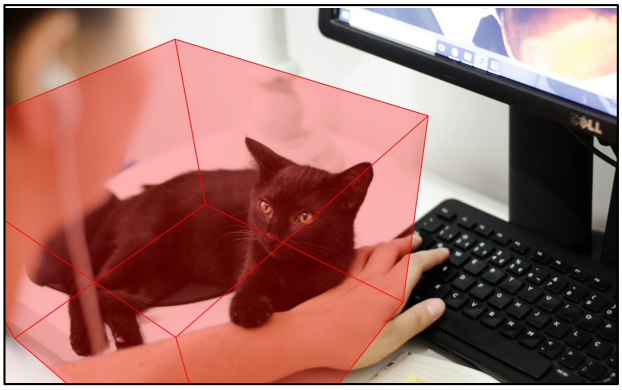
[16] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, 2020. 1

[17] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harness-

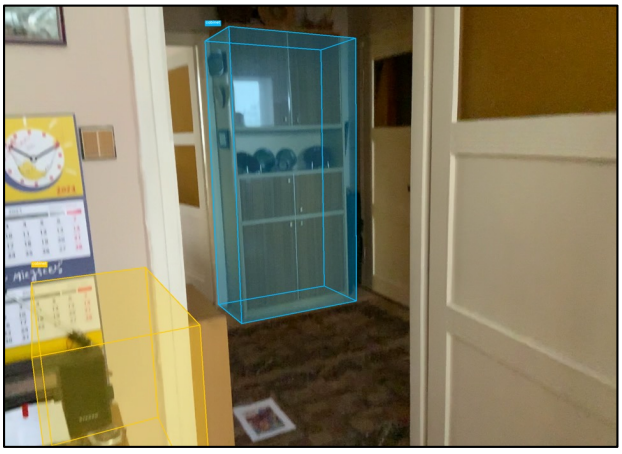
Orientation Error

Under-full Boxes

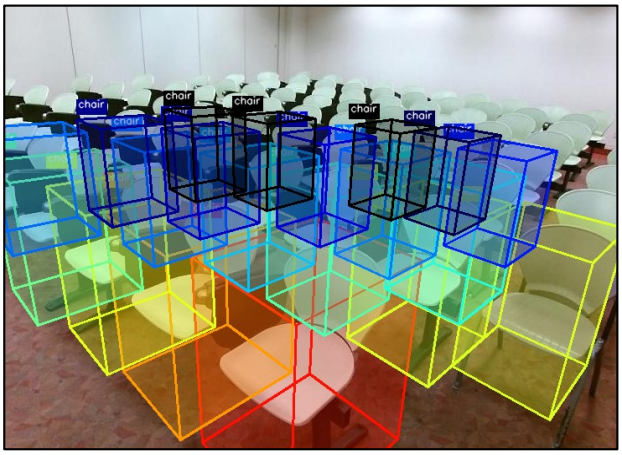
Location Mismatch



Depth Mismatch



False Positives



False Negatives

Figure B. **Visualization of failure cases**. We show several failure cases of our model. Due to the lack of diverse 3D annotations, similar to the baselines [139, 153], our model faces challenges when presented with scenes that exhibit very different focal length, spatial layouts, and textural details.

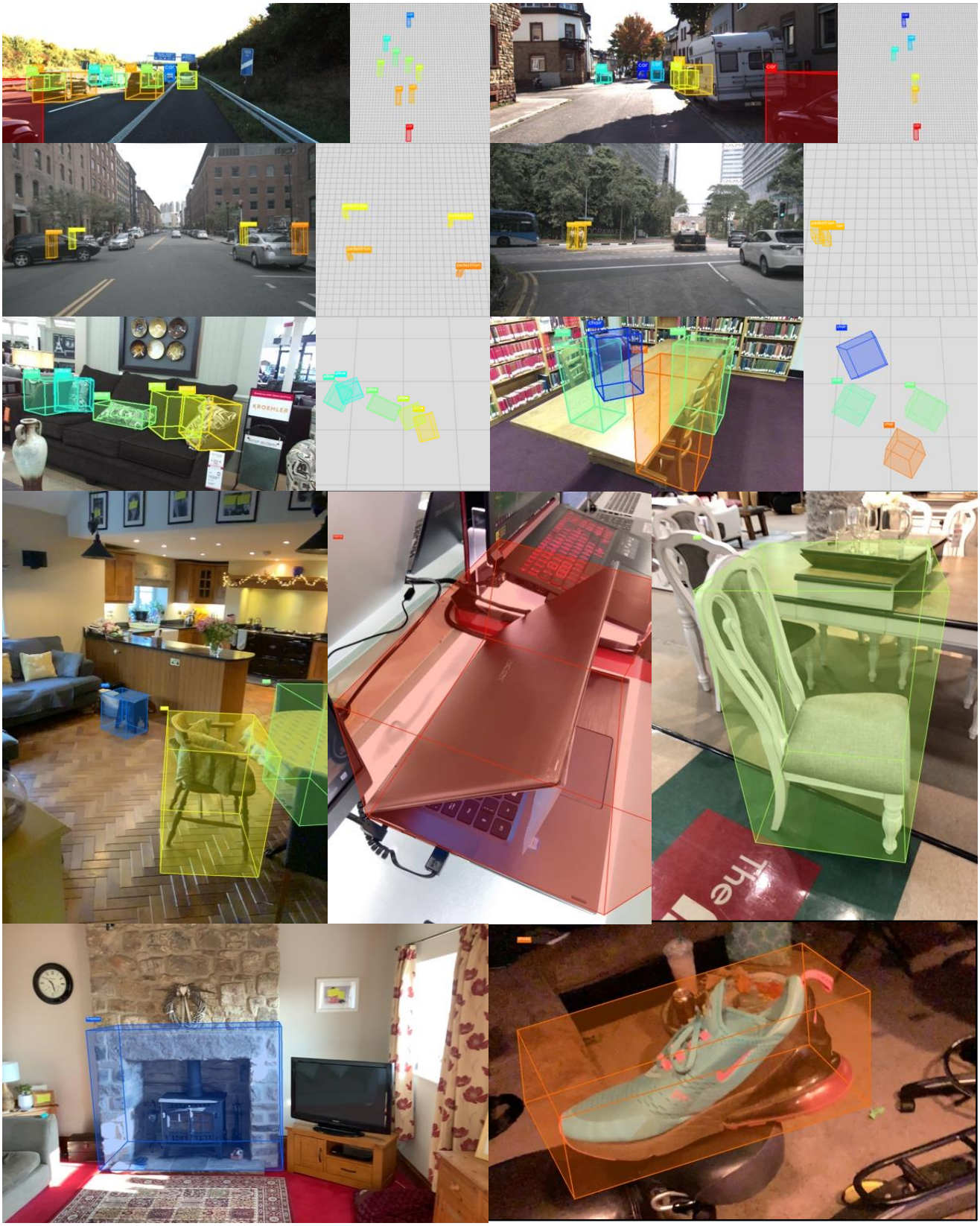


Figure C. Visualization of more indoor and outdoor successful cases.

ing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 4

[18] Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022. 4

[19] Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. Theoremqa: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901, 2023. 4

[20] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *ICDAR*, pages 1571–1576, 2019. 4

[21] Jang Hyun Cho, Boris Ivanovic, Yulong Cao, Edward Schmerling, Yue Wang, Xinshuo Weng, Boyi Li, Yurong You, Philipp Krähenbühl, Yan Wang, and Marco Pavone. Language-image models with 3d understanding. *arXiv preprint arXiv:2405.03685*, 2024. 1, 2

[22] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In *ACL*, pages 845–855, 2018. 4

[23] LLM-Red-Team Contributors. emo-visual-data: Emotion and visual data analysis project. https://github.com/LLM-Red-Team/emo-visual-data, 2024. 4

[24] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *NeurIPS*, 35:16344–16359, 2022. 3

[25] Markus Diem, Stefan Fiel, Florian Kleber, Robert Sablatnig, Jose M Saavedra, David Contreras, Juan Manuel Barrios, and Luiz S Oliveira. Icfhr 2014 competition on handwritten digit string recognition in challenging datasets (hdsrc 2014). In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 779–784. IEEE, 2014. 4

[26] Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. Vqa: A new dataset for real-world vqa on pdf documents. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 585–601. Springer, 2023. 4

[27] EmileEsmaili. sheet music clean ataset. https://huggingface.co/datasets/EmileEsmaili/sheet_music_clean, 2024. 4

[28] FastJobs. Visual emotional analysis dataset. https://huggingface.co/datasets/FastJobs/Visual_Emotional_Analysis, 2024. 4

[29] Manuel Faysse, Hugues Sibille, Tony Wu, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024. 4

[30] flytech. Python codes 25k dataset. https://huggingface.co/datasets/flytech/python-codes-25k, 2024. 4

[31] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017. 4

[32] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4

[33] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 4

[34] hazal karakus. mscoco-controlnet-canny-less-colors dataset. https://huggingface.co/datasets/hazal-karakus/mscoco-controlnet-canny-less-colors, 2024. 4

[35] Mengchao He, Yuliang Liu, Zhibo Yang, Sheng Zhang, Canjie Luo, Feiyu Gao, Qi Zheng, Yongpan Wang, Xin Zhang, and Lianwen Jin. Icpr2018 contest on robust reading for multi-type web images. In *2018 24th international conference on pattern recognition (ICPR)*, pages 7–12. IEEE, 2018. 4

[36] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 4

[37] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 4

[38] Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Victor Carbune, Jason Lin, Maria Wang, Srinivas Sunkara, Yun Zhu, and Jindong Chen. Screenqa: Large-scale question-answer pairs over mobile app screenshots. *arXiv preprint arXiv:2209.08199*, 2022. 4

[39] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar 2019 robust reading challenge on scanned receipts ocr and information extraction. In *International conference on document analysis recognition*, 2019. 4

[40] HugGAN. Wikiart dataset. https://huggingface.co/datasets/huggan/wikiart, 2024. 4

[41] iamtarun. Python code instructions 18k alpaca dataset. https://huggingface.co/datasets/iamtarun/python_code_instructions_18k_alpaca, 2024. 4

[42] ift. Handwriting forms dataset. https://huggingface.co/datasets/ift/handwriting_forms, 2024. 4

[43] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pages 1–6. IEEE, 2019. 4

[44] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*, 2018. 4

[45] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, pages 5648–5656, 2018. 4

[46] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. 4

[47] Kamizuru00. Diagram image to text dataset. https://huggingface.co/datasets/Kamizuru00/diagram_image_to_text, 2024. 4

[48] Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*, 2022. 4

[49] 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. 4

[50] 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. 4

[51] keremberke. Indoor scene classification dataset. https://huggingface.co/datasets/keremberke/indoor-scene-classification, 2024. 4

[52] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020. 4

[53] 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. 4

[54] Praveen Krishnan, Rama Kovvuri, Guan Pang, Boris Vassilev, and Tal Hassner. Textstylebrush: transfer of text aesthetics from a single example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9122–9134, 2023. 4

[55] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*, 2024. 4

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

[57] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 4

[58] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024. 4

[59] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 4

[60] Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset. *arXiv preprint arXiv:2403.09029*, 2024. 4

[61] Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024. 4

[62] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*, 2024. 4

[63] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *CVPR*, pages 14963–14973, 2023. 4

[64] Zekun Li, Yijun Lin, Yao-Yi Chiang, Jerod Weinman, Solenn Tual, Joseph Chazalon, Julien Perret, Bertrand Duménieu, and Nathalie Abadie. Icdar 2024 competition on historical map text detection, recognition, and linking. In *International Conference on Document Analysis and Recognition*, pages 363–380. Springer, 2024. 4

[65] W Lian, B Goodson, E Pentland, et al. Openorca: An open dataset of gpt augmented flan reasoning traces, 2023. 4

[66] 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. 4

[67] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022. 4

[68] 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. 4

[69] 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. 4

[70] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2023. 4

[71] Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, et al. Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity. *arXiv preprint arXiv:2407.15838*, 2024. 4

[72] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*, 2024. 4

[73] Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Icdar 2023 competition on hierarchical text detection and recognition. In *International Conference on Document Analysis and Recognition*, pages 483–497. Springer, 2023. 4

[74] LooksJuicy. Ruozhiba dataset. https://huggingface.co/datasets/LooksJuicy/ruozhiba, 2024. 4

[75] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021. 4

[76] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021. 4

[77] 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. 4

[78] 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. 4

[79] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022. 4

[80] Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*, 2024. 4

[81] Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. Weatherqa: Can multimodal language models reason about severe weather? *arXiv preprint arXiv:2406.11217*, 2024. 4

[82] Ibrahim Souleiman Mahamoud, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain d'Andecy, and Jean-Marc Ogier. Chic: Corporate document for visual question answering. In *International Conference on Document Analysis and Recognition*, pages 113–127. Springer, 2024. 4

[83] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 4

[84] U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5:39–46, 2002. 4

[85] 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. 4

[86] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*, 2023. 4

[87] Minesh Mathew, Lluis Gomez, Dimosthenis Karatzas, and CV Jawahar. Asking questions on handwritten document collections. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(3):235–249, 2021. 4

[88] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022. 4

[89] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *WACV*, pages 1527–1536, 2020. 4

[90] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*. BMVA, 2012. 4

[91] 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. 4

[92] Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*, 2024. 4

[93] Harold Mouchère, Christian Viard-Gaudin, Richard Zanibbi, and Utpal Garain. Icfhr2016 crohme: Competition on recognition of online handwritten mathematical expressions. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 607–612. IEEE, 2016. 4

[94] mPLUG. Docreason25k dataset. https://huggingface.co/datasets/mPLUG/DocReason25K, 2024. 4

[95] mychen76. Invoices and receipts ocr v1 dataset. https://huggingface.co/datasets/mychen76/invoices-and-receipts_ocr_v1, 2024. 4

[96] Abhilash Nandy, Yash Agarwal, Ashish Patwa, Millon Madhur Das, Aman Bansal, Ankit Raj, Pawan Goyal, and Niloy Ganguly. Yesbut: A high-quality annotated multimodal dataset for evaluating satire comprehension capability of vision-language models. *arXiv preprint arXiv:2409.13592*, 2024. 4

[97] OleehyO. Latex formulas dataset. https://huggingface.co/datasets/OleehyO/latex-formulas, 2024. 4

[98] OpenGVLab. Sharegpt-4o dataset. https://huggingface.co/datasets/OpenGVLab/ShareGPT-4o, 2024. 4

[99] parasam. Captcha dataset. https://www.kaggle.com/datasets/parsasam/captcha-dataset, 2024. 4

[100] Renjie Pi, Jianshu Zhang, Jipeng Zhang, Rui Pan, Zhekai Chen, and Tong Zhang. Image textualization: An automatic framework for creating accurate and detailed image descriptions. *arXiv preprint arXiv:2406.07502*, 2024. 4

[101] PixArt-alpha. Sam-llava-captions10m dataset. https://huggingface.co/datasets/PixArt-alpha/SAM-LLaVA-Captions10M, 2024. 4

[102] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 4

[103] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer, 2020. 4

[104] Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava DaS, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. Memotion 2: Dataset on sentiment and emotion analysis of memes. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*, 2022. 4

[105] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28, 2015. 4

[106] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, pages 146–162, 2022. 4

[107] Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1466–1476, 2015. 4

[108] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *AAAI*, pages 8876–8884, 2019. 4

[109] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439, 2019. 4

[110] shreyanshu09. Block diagram dataset. https://huggingface.co/datasets/shreyanshu09/Block_Diagram, 2024. 4

[111] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023. 4

[112] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 4

[113] skvarre. Movie posters-100k dataset. https://huggingface.co/datasets/skvarre/movie_posters-100k, 2024. 4

[114] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 1

[115] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017. 4

[116] Hamed Rahimi Sujet AI, Allaa Boutaleb. Sujet-finance-qa-vision-100k: A large-scale dataset for financial document vqa, 2024. 4

[117] Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *ICDAR*, pages 1557–1562, 2019. 4

[118] TAL. Tal open dataset. https://ai.100tal.com/dataset, 2023. 4

[119] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13636–13645, 2023. 4

[120] Benny J Tang, Angie Boggust, and Arvind Satyanarayan. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*, 2023. 4

[121] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*, 2024. 4

[122] Qwen Team. Qwen2.5: A party of foundation models, 2024. 3

[123] Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants. https://huggingface.co/datasets/teknium/OpenHermes-2.5, 2023. 4

[124] TIGER-Lab. Visualwebinstruct dataset. https://huggingface.co/datasets/TIGER-Lab/VisualWebInstruct, 2024. 4

[125] George Tom, Minesh Mathew, Sergi Garcia-Bordils, Dimosthenis Karatzas, and CV Jawahar. Icdar 2023 competition on roadtext video text detection, tracking and recognition. In *International Conference on Document Analysis and Recognition*, pages 577–586. Springer, 2023. 4

[126] Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*, 2023. 4

[127] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark

for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 4

[128] VQAonDB. Vqaondb dataset. https://ilocr.iiit.ac.in/vqabd/. 4

[129] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023. 4

[130] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *CVPR*, pages 10126–10135, 2020. 4

[131] wendlerc. Renderedtext dataset. https://huggingface.co/datasets/wendlerc/RenderedText, 2024. 4

[132] Tobias Weyand, André Araujo, Bingyi Cao, and Jack Sim. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *CVPR*, 2020. 4

[133] Siwei Wu, Kang Zhu, Yu Bai, Yiming Liang, Yizhi Li, Haoning Wu, Jiaheng Liu, Ruibo Liu, Xingwei Qu, Xuxin Cheng, et al. Mmra: A benchmark for multi-granularity multi-image relational association. *arXiv preprint arXiv:2407.17379*, 2024. 4

[134] Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. Toward understanding wordart: Corner-guided transformer for scene text recognition. In *European conference on computer vision*, pages 303–321. Springer, 2022. 4

[135] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*, 2024. 4

[136] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023. 4

[137] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint*, 2024. 4

[138] Kaiyu Yang, Olga Russakovsky, and Jia Deng. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2051–2060, 2019. 4

[139] Jin Yao, Hao Gu, Xuweiyi Chen, Jiayun Wang, and Zezhou Cheng. Open vocabulary monocular 3d object detection. *arXiv preprint arXiv:2411.16833*, 2024. 5, 6

[140] 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. 4

[141] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[142] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 4

[143] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023. 4

[144] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024. 4

[145] Wenwen Yu, Chengquan Zhang, Haoyu Cao, Wei Hua, Bohan Li, Huang Chen, Mingyu Liu, Mingrui Chen, Jianfeng Kuang, Mengjun Cheng, et al. Icdar 2023 competition on structured text extraction from visually-rich document images. In *International Conference on Document Analysis and Recognition*, pages 536–552. Springer, 2023. 4

[146] Youngjoon Yu, Sangyun Chung, Byung-Kwan Lee, and Yong Man Ro. Spark: Multi-vision sensor perception and reasoning benchmark for large-scale vision-language models. *arXiv preprint arXiv:2408.12114*, 2024. 4

[147] Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, et al. Advancing llm reasoning generalists with preference trees. *arXiv preprint arXiv:2404.02078*, 2024. 4

[148] Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large chinese text dataset in the wild. *Journal of Computer Science and Technology*, 34: 509–521, 2019. 4

[149] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023. 4

[150] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 3

[151] Bo-Wen Zhang, Yan Yan, Lin Li, and Guang Liu. Infinitymath: A scalable instruction tuning dataset in programmatic mathematical reasoning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5405–5409, 2024. 4

[152] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327, 2019. 4

[153] Hanxue Zhang, Haoran Jiang, Qingsong Yao, Yanan Sun, Renrui Zhang, Hao Zhao, Hongyang Li, Hongzi Zhu, and

Zetong Yang. Detect anything 3d in the wild. In *ICCV*, 2025. 1, 5, 6

[154] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *ICDAR*, pages 1577–1581, 2019. 4

[155] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*, 2024. 4

[156] Xiaokang Zhang, Jing Zhang, Zeyao Ma, Yang Li, Bohan Zhang, Guanlin Li, Zijun Yao, Kangli Xu, Jinchang Zhou, Daniel Zhang-Li, et al. Tablellm: Enabling tabular data manipulation by llms in real office usage scenarios. *arXiv preprint arXiv:2403.19318*, 2024. 4

[157] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 4

[158] Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. Multihiertt: Numerical reasoning over multi hierarchical tabular and textual data. *arXiv preprint arXiv:2206.01347*, 2022. 4

[159] Yilun Zhao, Chen Zhao, Linyong Nan, Zhenting Qi, Wenlin Zhang, Xiangru Tang, Boyu Mi, and Dragomir Radev. Robut: A systematic study of table qa robustness against human-annotated adversarial perturbations. *arXiv preprint arXiv:2306.14321*, 2023. 4

[160] Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhu Chen, and Xiang Yue. Opencodeinterpreter: Integrating code generation with execution and refinement. *arXiv preprint arXiv:2402.14658*, 2024. 4

[161] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 4

[162] Chenchen Zhu, Fanyi Xiao, Andrés Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Chang, Roshan Sumbaly, and Zhicheng Yan. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 4

[163] 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. 4