LocateAnything: Fast and High-Quality Vision-Language Grounding with Parallel Box Decoding

Shihao Wang[1*], **Shilong Liu**[2*], **Yuanguo Kuang**[1], **Xinyu Wei**[1], **Yangzhou Liu**[3], **Zhiqi Li**, **Yunze Man**[4*], **Guo Chen**[3*], **Andrew Tao**, **Guilin Liu**, **Jan Kautz**, **Lei Zhang**[1], **Zhiding Yu**[†]

Links: GitHub | HF Model | HF Demo | Project Page

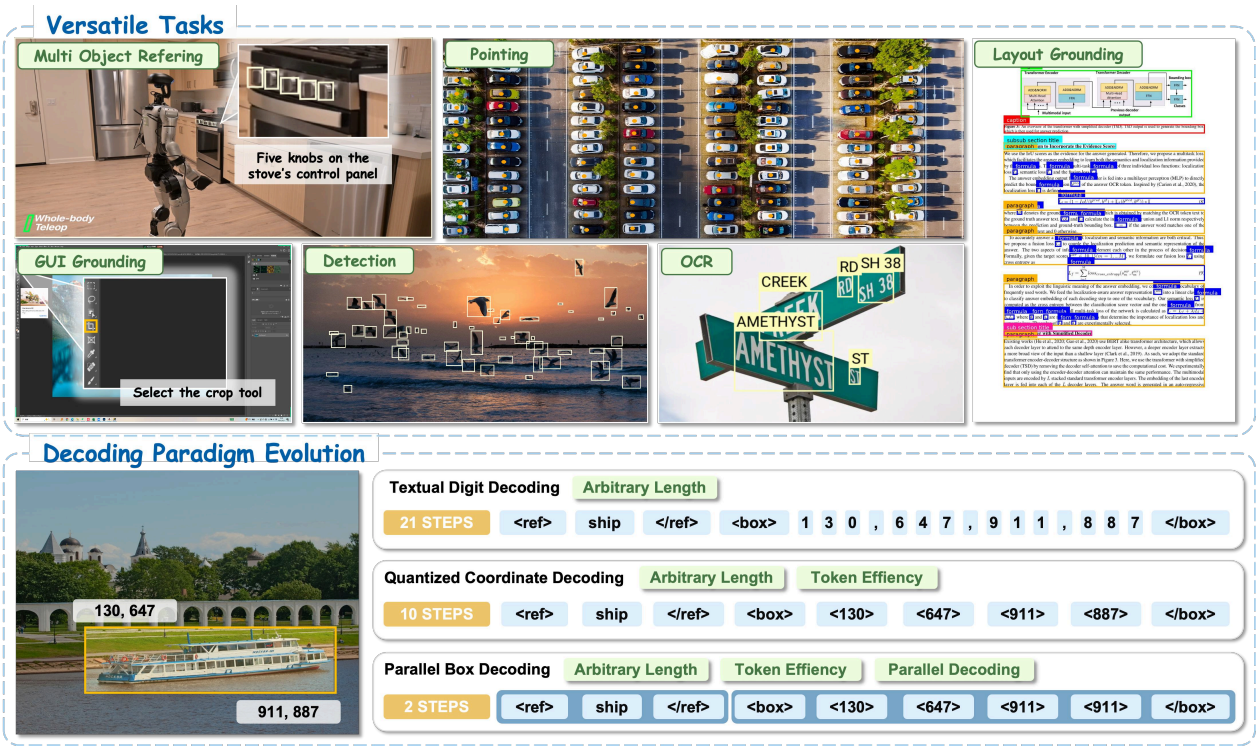


Figure 1: **Versatile tasks of LocateAnything with parallel box decoding. Top:** LocateAnything supports diverse localization tasks under a unified vision-language model. **Bottom:** Textual digit decoding spells coordinates digit by digit, and quantized coordinate decoding predicts coordinate tokens sequentially. In contrast, **Parallel Box Decoding** predicts each geometric unit (*e.g.*, a bounding box) in a single forward pass.

Abstract

Vision-language models (VLMs) commonly formulate visual grounding and detection as a coordinate-token generation problem, serializing each 2D box into multiple 1D tokens that are learned and decoded largely independently. This token-by-token decoding mismatches the coupled structure of box geometry and creates a practical inference bottleneck due to strictly sequential generation. We introduce LocateAnything, a unified generative grounding and detection framework based on Parallel Box Decoding (PBD). By decoding geometric elements such as bounding boxes and points as atomic units in a single step, LocateAnything preserves intra-box geometric coherence and unlocks substantial parallelism. We show that PBD improves both decoding throughput and localization accuracy. We further develop a scalable data engine and curate LocateAnything-Data, a large-scale dataset with more than 138 million training samples, substantially increasing data diversity for high-precision localization. Extensive evaluations show that LocateAnything advances the speed–accuracy frontier, achieving significantly higher decoding throughput while improving high-IoU localization quality across diverse benchmarks. The results highlight the complementary benefits of Parallel Box Decoding and large-scale training data in enabling efficient and precise unified visual grounding and detection.

* *Work done during an internship at NVIDIA.* † *Corresponding author: zhidingy@nvidia.com.*
Additional affiliations: [1] *The Hong Kong Polytechnic University,* [2] *Princeton University,* [3] *Nanjing University,* [4] *University of Illinois Urbana-Champaign.*

1. Introduction

Vision-language models (VLMs) (Bai et al., 2025b; Chen et al., 2025; Deshmukh et al., 2025; Huang et al., 2026; Wang et al., 2025a; Yang et al., 2025a) are increasingly adopted as a general-purpose backbone for interactive and embodied systems due to their broader knowledge and stronger instruction-following capabilities than conventional specialized models (Carion et al., 2020; Liu et al., 2023d; Ren et al., 2016; Zhang et al., 2022). To act in the world, VLMs (Azzolini et al., 2025; Bai et al., 2025b; Fu et al., 2025b; Wang et al., 2025a; Zhan et al., 2024) must be tightly grounded in *perception* — in particular, they *localize* task-relevant entities (*e.g.*, objects (Jiang et al., 2025a; Wang et al., 2023b; Yu et al., 2025; Zhang et al., 2024b), UI elements (Feizi et al., 2025b; Lin et al., 2024a; Liu et al., 2025e; Nayak et al., 2025), regions (Cheng et al., 2024; Heinrich et al., 2025; Lai et al., 2024a; Ranzinger et al., 2024; Ren et al., 2024b; Yuan et al., 2025)) from natural-language intents with high quality and low latency, which requires high vision-language grounding capabilities.

Object detection and grounding in VLMs (Jiang et al., 2025a; Li et al., 2025a; Man et al., 2025; Peng et al., 2023; Yu et al., 2025; Zhan et al., 2024; Zhang et al., 2024c) are often formulated as a *generative* problem. Under the next-token prediction (NTP) paradigm (Chen et al., 2022b; Jiang et al., 2025a; Peng et al., 2023), a VLM can answer open-ended queries by emitting spatial coordinates as a token sequence. As illustrated in the bottom panel of Fig. 1, existing methods (Jiang et al., 2025a; Peng et al., 2023; Qi et al., 2025; You et al., 2024; Zhang et al., 2024c) commonly represent coordinates as either **Textual Digits** (*e.g.*, "1024" as "1", "0", "2", "4") or **Quantized Tokens** (*e.g.*, $x_1 \rightarrow y_1 \rightarrow x_2 \rightarrow y_2$). Despite their differences, these representations serialize a 2D geometric object into a 1D stream, forcing token-by-token generation at inference time. This token-level sequential decoding becomes a practical bottleneck (higher latency and lower throughput) and under-utilizes the strong structured correlation among coordinates (x_1, y_1, x_2, y_2).

Multi-Token Prediction (MTP) (Liu et al., 2025c; Nie et al., 2025; Ye et al., 2025b; Zeng et al., 2025) offers a natural approach to reducing decoding steps by predicting multiple tokens in parallel. In language modeling, MTP is usually implemented by randomly (i) choosing positions in the sequence and training the model to predict a following span in parallel (*i.e.*, next-block prediction) (Cai et al., 2024b; Li et al., 2025c; Liu et al., 2024a, 2025c), or (ii) masking some tokens of the sequence and training the model to reconstruct the original text, such as masked diffusion modeling (Arriola et al., 2025; Li et al., 2022; Liu et al., 2025b; Nie et al., 2025). However, these formulations are largely *structure-agnostic*: they treat inputs as generic token streams and mainly capture correlations driven by co-occurrence. Inferring the missing tokens from random subsets requires the model to represent complex and irregular conditional distributions. For tightly coupled units such as bounding boxes, this supervision does not match well the training objective because it can learn to generate token combinations across bounding-box boundaries and even object categories, as demonstrated in Fig. 2. Consequently, the model must fit many unreliable patterns, inducing spurious correlations, sacrificing structured decoding, and amplifying error propagation, which together reduce accuracy, reliability, and decoding speed.

To reconcile high-throughput decoding with reliable localization, we propose **LocateAnything**, a unified framework for VLM-based visual detection and grounding built upon **Parallel Box Decoding (PBD)**. Our key idea is to align MTP blocks with structured units: during training, LocateAnything treats each bounding box (or point) as an *atomic unit* and learns to predict the full coordinate set (x_1, y_1, x_2, y_2) in one parallel step. This *box-aligned* training target avoids arbitrary chunking of coordinate tokens. As a result, our strategy improves the localization performance of the model, while simultaneously unlocking the speed benefits of parallel decoding.

With the proposed PBD, we study various strategies for structured bounding-box decoding to balance throughput and accuracy. Our observations motivate a flexible inference design to meet different latency–robustness requirements by providing three on-demand modes. (i) **Fast Mode** (MTP) predicts full boxes in parallel for maximum throughput, which is suitable for latency- and compute-constrained settings, such as on-device robotics and embodied agents. (ii) **Slow Mode** (NTP) decodes coordinate tokens autoregressively for maximum

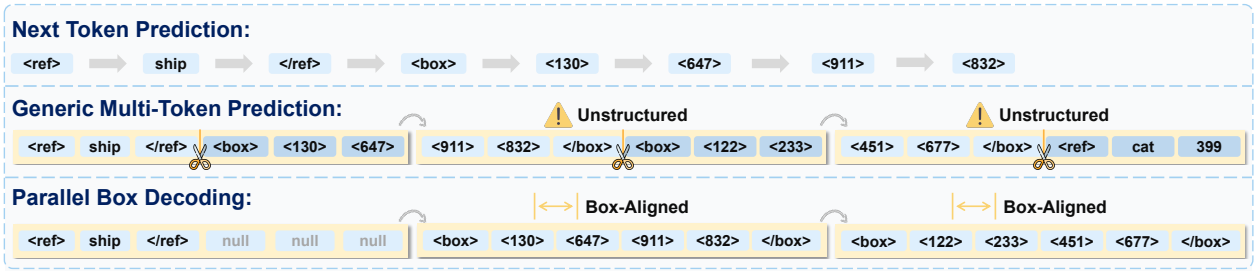


Figure 2: **Comparison of Token Decoding Methods.** The NTP generates coordinate values one-by-one. The standard MTP method results in irregular distributions and non-coherent, unstructured patterns. Our proposed PBD generates a single atomic box (or point) unit in a parallel step, ensuring box-aligned and structured output.

stability, which is appropriate for high-precision labeling, final-pass dataset curation, and accuracy-oriented offline evaluation. (iii) **Hybrid Mode** uses Fast Mode by default and falls back to Slow Mode when the parallel output is unreliable, *e.g.*, due to format or consistency violations; this mode is intended for production pipelines that require both speed and accuracy. Overall, Hybrid Mode preserves most of the speed gains of parallel decoding while maintaining robust outputs.

Our main contributions are summarized as follows:

- We introduce **LocateAnything**, an early exploration of applying multi-token prediction to VLM-based detection/grounding via **Parallel Box Decoding**, performing box-aligned decoding to improve throughput and accuracy.
- We present a Hybrid decoding policy that detects unreliable parallel blocks and performs localized NTP re-decoding only for the problematic block, reducing worst-case failures while retaining most speed gains.
- Extensive evaluations, including layout grounding, long-tail detection, and GUI grounding, show that LocateAnything advances the **speed–accuracy frontier**, outperforming the SOTA by a large margin. It achieves up to **2.5**$\times$ higher decoding throughput while improving localization quality.

2. Related Work

Visual Detection and Grounding in VLMs. Visual grounding/detection tasks traditionally rely on task-specific heads (Carion et al., 2020; Jiang et al., 2024; Ren et al., 2016, 2024a), but recent VLMs like Qwen-VL series (Bai et al., 2025a,b), InternVL (Chen et al., 2023c) and Shikra (Chen et al., 2023a) formulate it as an autoregressive token generation problem. This generative paradigm, however, often suffers from structural hallucinations and high latency (Li et al., 2023a). To mitigate these issues, Rex-Omni (Jiang et al., 2025a) employs point-based prediction, while Patch-as-Decodable-Token (PaDT) (Su et al., 2026) and Groma (Ma et al., 2024b) utilize visual reference tokens to point directly to image patches. Complementary innovations such as Pink (Xuan et al., 2024), ViP-LLaVA (Cai et al., 2024a), Griffon (Zhan et al., 2024), DnU (Lin et al., 2024b) and PAM (Lin et al., 2025) focus on enhancing 2D referential comprehension through visual prompt engineering and multi-granularity feature scaling. LLMDet (Fu et al., 2025b) boosts detection recall by data distribution tuning. To bypass serial decoding bottlenecks, WeDetect (Fu et al., 2025a) treats detection as a parallel retrieval task. Advanced perception logic is further integrated via Chain-of-Thought (CoT) (Qi et al., 2025), while post-training strategies such as Vision-R1 (Zhan et al., 2025), UniVG-R1 (Bai et al., 2025c) and GW-VLM (Jiang et al., 2026) utilize reinforcement learning to align model outputs with visual feedback and reduce grounding errors (Zhang et al., 2024c).

Parallel Decoding via MTP and Diffusion LLMs. To mitigate autoregressive latency, parallel generation techniques such as MTP (Cai et al., 2024b; Gloeckle et al., 2024; Samragh et al., 2025) predict multiple future tokens simultaneously, often coupled with speculative decoding to accelerate inference. Recent extensions such as Future Summary Prediction (Mahajan et al., 2025) capture long-term dependencies via auxiliary heads. Concurrently, Diffusion Language Models (DLMs) such as LLaDA (Nie et al., 2025), Dream (Ye et al., 2025b), and

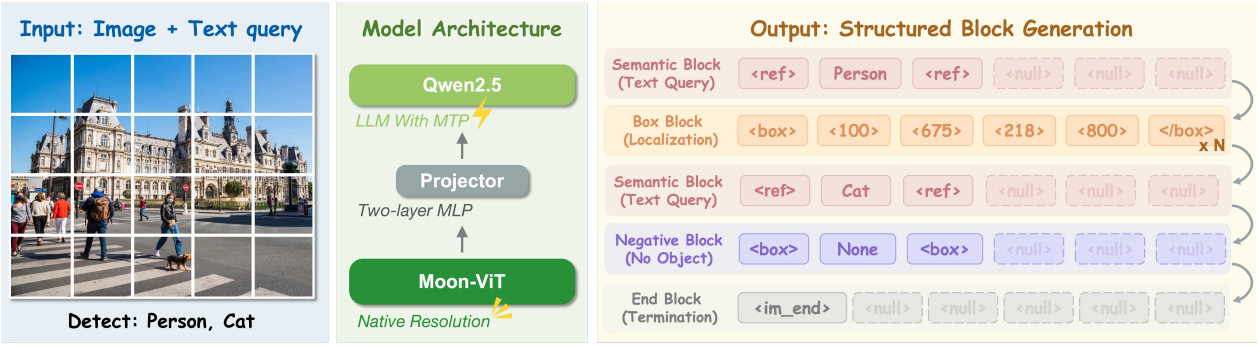


Figure 3: **Architecture and Block-Based Output Representation.** LocateAnything formulates localization as generating a sequence of fixed-length, *box-aligned atomic blocks*. Four functional block types—Semantic, Box, Negative, and End blocks—are defined to jointly specify predicted entities or termination states.

DiffuCoder (Gong et al., 2025) frame sequence generation as a discrete denoising process, enabling bidirectional context modeling and non-autoregressive decoding. Hybrid semi-autoregressive paradigms, including Block Diffusion (Arriola et al., 2025), SDLM (Liu et al., 2025c) and Fast-dLLM v2 (Wu et al., 2025a), decode fixed-size token blocks in parallel while maintaining causal dependencies to preserve KV-caching compatibility. More advanced frameworks (Lu et al., 2025a; Wang et al., 2025b) unlock inter-block parallelism and adaptive block scheduling. These paradigms have been extended to the multimodal domain via DiffusionVL (Zeng et al., 2025), translating autoregressive LMMs into high-performance diffusion-based models.

LocateAnything differs from existing works in two key aspects. First, instead of generating bounding boxes via slow NTP, we output the complete box in a single parallel step. Second, recent MTP paradigms group tokens into arbitrary chunks. Instead, our PBD treats the entire coordinate set as a single atomic block, resolving both the fragmentation of NTP and the arbitrary chunking of MTP, seamlessly unifying high throughput with structural coherence.

3. Method

This section presents **LocateAnything**, a fast and effective framework that integrates **Parallel Box Decoding (PBD)** into VLMs for visual detection and grounding. Section 3.1 introduces the model architecture and the block-based output formulation. Section 3.2 details the joint training strategy, which aligns NTP with block-level MTP. Section 3.3 describes the on-demand inference mechanism, featuring a hybrid mode that dynamically balances decoding throughput and robustness. Finally, Section 3.4 outlines the construction of our large-scale training dataset, **LocateAnything-Data**.

3.1. Model Architecture and Formulation

Overview. As illustrated in Fig. 3, LocateAnything builds upon a native-resolution VLM pre-trained on large-scale image-text corpora. The architecture comprises a Moon-ViT (Kimi Team, 2025) vision encoder and a Qwen2.5 (Qwen Team, 2024) language decoder, bridged by a MLP projector. Given an input image $\mathcal{I}$, the vision encoder extracts visual tokens $Z = \text{Encoder}(\mathcal{I})$ at the native resolution, preserving the fine-grained spatial details crucial for high-precision localization. These tokens are subsequently fed into the language model, which directly converts them into a sequence of box-aligned block-level predictions.

Block-Based Output Formulation. To facilitate PBD, we abandon standard NTP coordinate generation. Instead, continuous coordinates are normalized to $[0, 1000]$, discretized into tokens (Chen et al., 2022b; Jiang et al., 2025a), and reorganized into a sequence of blocks $\mathbf{B} = (b_1, b_2, \ldots, b_N)$. Conditioned on the visual features Z and a text query $\mathcal{E}$, the joint probability is formulated as $P(\mathbf{B} \mid \mathcal{Z}, \mathcal{E}) = \prod_{i=1}^{N} P(b_i \mid b_{<i}, Z, \mathcal{E})$.

Each block b_i acts as an atomic unit of constant length $L = 6$, accommodating a bounding box and two structural tokens (*e.g.*, `<box>` and `</box>`). To guarantee uniform tensor shapes for parallel decoding, any

unoccupied positions are padded with a <null> token. As depicted in Fig. 3, we define four functional block types. *(1) Semantic Block*: Encodes the linguistic identity. If an expression exceeds the capacity of a single block, it is partitioned across multiple consecutive blocks. *(2) Box Block*: Uses four quantized coordinates representing the bounding boxes. *(3) Negative Block*: Explicitly indicates the absence of a queried object. *(4) End Block*: Signals the termination of the generation process.

3.2. Training Design

Our method treats bounding box coordinates as an indivisible atomic unit, enforcing structured supervision and unlocking the capability for parallel generation. However, parallelizing the output directly in the training phase risks disrupting the model's inherent causal reasoning process. To resolve this issue, we introduce a dual-formulation training strategy that jointly optimizes two aligned representations: the NTP sequence to preserve the causal reasoning ability, and the block-wise MTP formulation for box-aligned predictions. To implement this, a single concatenated input sequence is constructed: $x_{\text{all}} = x_{\text{vis}} \oplus x_{\text{q}} \oplus x_{\text{ntp}} \oplus x_{\text{blk}}$, where $\oplus$ denotes sequence concatenation. The terms x_{vis} and x_{q} serve as the shared context (visual and text query inputs), x_{ntp} represents the standard NTP input sequence, and x_{blk} is the block-wise MTP input sequence. Essentially, they represent the identical ground truth in two distinct formats: a *token-level representation* and a *block-level representation*.

Specifically, inspired by (Liu et al., 2025a,c), x_{blk} is constructed by traversing x_{ntp} from left to right, splitting and padding the sequence according to our previously defined block rules. Within each block, we retain the first token to serve as the prediction context, while replacing all subsequent tokens with [mask] tokens. This structure prompts the model to simultaneously predict all masked tokens within the block in a single cohesive step. Notably, if the block size is set to 1, this MTP formulation naturally becomes equivalent to standard NTP.

Attention Mask Design. The core challenge of this dual-sequence formulation is how to isolate the NTP and MTP streams while allowing both to leverage the shared context. This is achieved through a specialized attention mask (as shown in Fig. 4), which dictates information flow via three distinct behaviors:

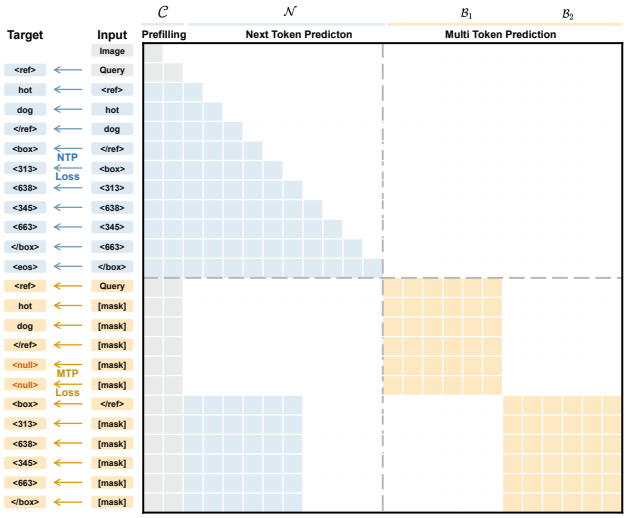


Figure 4: **Architecture and Block-Based Output Representation.** LocateAnything formulates localization as generating a sequence of fixed-length, *box-aligned atomic blocks*. Four functional block types—Semantic, Box, Negative, and End blocks—are defined to jointly specify predicted entities or termination states.

Causal Attention for NTP. To preserve the original language capabilities of the VLM, the shared context (x_{vis} and x_{q}) and the NTP sequence (x_{ntp}) collectively employ a causal attention mask. Tokens within these segments can only attend to preceding tokens. Crucially, they are restricted from attending to x_{blk} to prevent data leakage. This strict causal formulation perfectly aligns with the standard KV Cache usage during inference.

Causal Flow Across Blocks. To align with the semi-autoregressive generation process, attention across different blocks in x_{blk} is strictly causal. Tokens in the active block can attend to the shared context and all previously committed blocks, but cannot see future blocks. This historical visibility enables the model to learn dependencies between different box predictions, effectively mitigating duplicate or missing bounding boxes.

Bidirectional Intra-Block Attention. Following the block-causal design widely adopted in recent generative modeling (Arriola et al., 2025; Fu et al., 2025c; Nie et al., 2025; Wang et al., 2025b; Wu et al., 2025a,b),

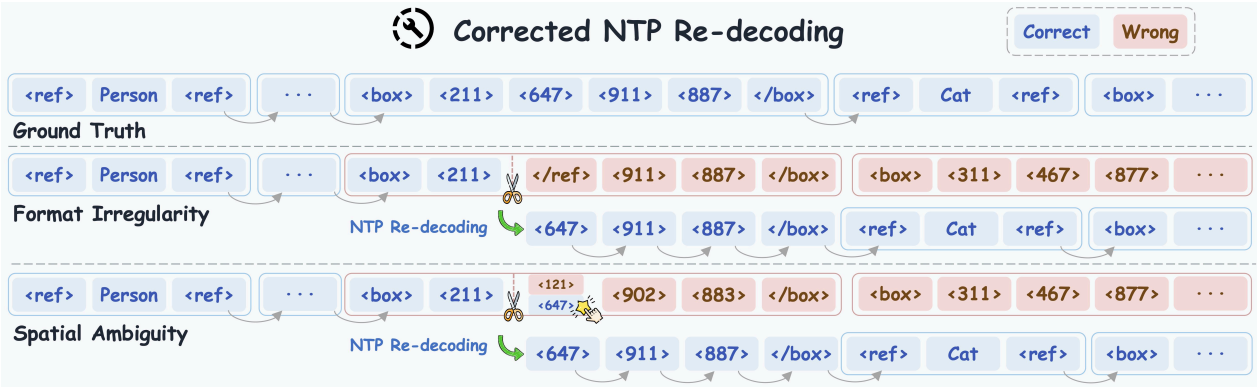


Figure 5: **Corrected NTP Re-decoding.** When parallel decoding encounters *Format Irregularity* or *Spatial Ambiguity*, the model discards the erroneous block and reverts to standard NTP to ensure robust predictions.

tokens within the same block share bidirectional attention. This fully-connected intra-block interaction allows the model to capture complex internal relationships (*e.g.*, geometric dependencies among a set of coordinates) and resolve all internal tokens simultaneously within a single functional unit.

Objective. Guided by this mask, we jointly minimize the cross-entropy losses for both sequences, *i.e.*, $\mathcal{L} = \mathcal{L}_{\text{ntp}} + \mathcal{L}_{\text{blk}}$.

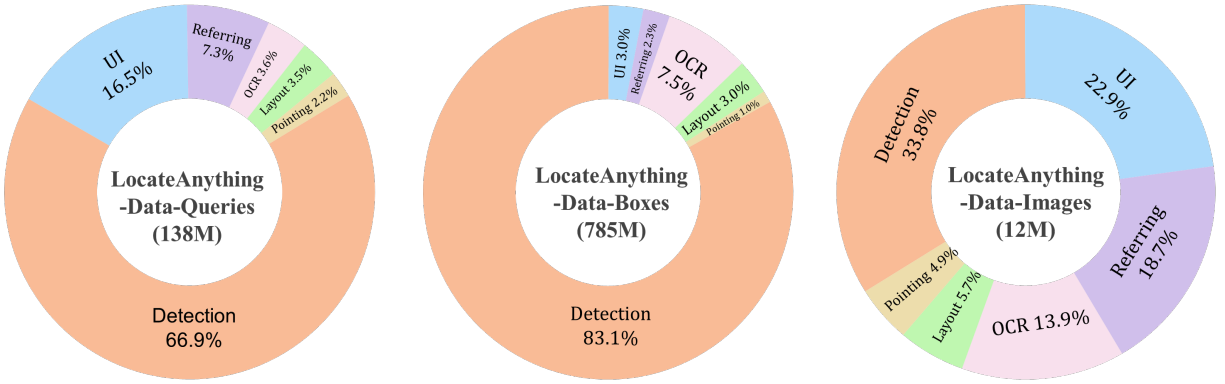
3.3. On-Demand Inference Modes

While our proposed PBD significantly accelerates inference, parallel decoding faces an inherent exploration-exploitation dilemma in highly complex scenes, as shown in Fig. 5. The first is *Format Irregularity*, which occurs in complex scenes containing multiple instances across categories. During parallel decoding, the model may struggle at category boundaries, hesitating between continuing to predict for the current class or transitioning to a new class. This uncertainty manifests as malformed syntax within a single predicted block, erroneously mixing structural and coordinate tokens (*e.g.*, `<box><211></ref><911><887></box>`). The second is *Spatial Ambiguity*, which arises when objects are densely arranged in regular grids, such as rows or columns. The MTP approach can blur spatial boundaries and output an intermediate coordinate situated between two objects, consequently producing low IoU predictions.

Both failure patterns can be effectively resolved using an NTP fallback mechanism. The NTP prediction can achieve higher precision when handling complex category transitions and dense spatial layouts. Therefore, during MTP inference, we continuously validate the syntactic integrity and monitor spatial confidence. Specifically, an ambiguity trigger is activated if two conditions are met simultaneously: (1) the top-1 coordinate token's probability is below 0.7, and (2) the max-min difference among the top-5 coordinate tokens exceeds 80 within the [0, 1000] normalized space. Upon detection of a format violation or high spatial ambiguity, the compromised block is discarded, and the generation reverts to the last verified prefix. NTP is then employed to autoregressively generate the tokens for the specific problematic block. Once the block is completed, the model seamlessly switches back to MTP for subsequent predictions.

Based on the above discussion, we propose three on-demand inference modes to balance throughput and spatial robustness. *(1) Slow Mode*, which generates the output token-by-token using standard NTP. *(2) Fast Mode*, which leverages MTP to predict box-aligned blocks. For each block, `<null>` padding tokens are discarded, and the remaining tokens are appended to the output; the committed tokens are stored in the key-value cache and serve as causal context for subsequent prediction steps. *(3) Hybrid Mode*, which employs MTP by default but seamlessly switches to NTP when parallel outputs become unreliable.

Inference-Time Attention Mask. During inference, the attention mask for each MTP decoding step mirrors the training-time block-causal pattern illustrated in Fig. 4. All previously committed tokens in the KV cache follow



Detection (93M, 66.9%)	Object365 (47.3M)	OpenImages (41.0M)	V3Det (2.7M)	EgoObjects (473.4K)	BDD100K (420.0K)
NuImages (382.1K)	PACO (368.7K)	DeepFashion2 (312.2K)	PartImageNet++ (269.4K)	SKU110K (147.7K)	CrowdHuman (15.0K)
MOT17DET (5.3K)	MOT20DET (8.9K)	OWDOD (8.9K)			
UI (23M, 16.5%)	OSAtlas (14.3M)	GroundCUA (5.5M)	ScaleCUA (1.6M)	GTAGrounding (1.5M)	MultiUI (5.8K)
Referring (10M, 7.3%)	OpenImages (6.2M)	Object365 (1.2M)	UnSplash (1.2M)	Flickr30kEntities (665.2K)	gRefCOCO (209.3K)
RoboAfford (200.4K)	RefCOCO (120.6K)	RefCOCO+ (120.1K)	RefCOCOg (80.5K)	HumanPart (44.2K)	HumanRef (37.1K)
OCR (5M ,3.6%)	BLIP3OCR (3.8M)	IDLOCR (1.1M)	Art (55.3K)	LSVT (30.0K)	TextOCR (21.8K)
ReCTS (20.0K)	HierText (8.3K)	RCTW (8.0K)	COCO_Text V2.0 (1.5K)	WildReceipt (1.3K)	ICDAR2015 (1.0K)
SROIE (0.7K)	ICDAR2013 (0.2K)				
Layout (5M, 3.5%)	PubLayNet (3.4M)	DocLayNet (1.0M)	TableBank (278.6K)	M6Doc (81.7K)	CDLA (70.0K)
TabRecSet (32.5K)					
Pointing (3M, 2.2%)	Object365 (426.3K)	OpenImages (435.6K)	PixmoPoints (2.1M)	RoboAfford (161.7K)	

Figure 6: Overview of the **LocateAnything-Data** dataset. The pie charts illustrate the task distribution across natural language queries, bounding boxes, and unique images. The bottom panel provides a detailed breakdown specifically for the language queries, showing the absolute count and percentage for each task category.

standard causal attention, while the n_{future} tokens in the current MTP block attend to each other bidirectionally, enabling parallel token prediction. Meanwhile, the current block can attend to all preceding blocks but is prevented from accessing subsequent ones. After each MTP step, the KV cache is truncated to retain only committed tokens, evicting mask tokens and the duplicated anchor to ensure the cache stays consistent with the causal prefix seen during training.

3.4. LocateAnything-Data

To train a highly capable model for general-purpose visual detection and grounding, we curate **LocateAnything-Data**, a large-scale, multi-domain dataset. The dataset construction details can be found in the **supplementary**.

As illustrated in Fig. 6, the dataset contains 12M unique images and 138M natural language queries. Furthermore, the dataset includes 785M annotated bounding boxes, providing massive and dense supervisory signals to guide the spatial learning of the LocateAnything model. The training corpus is categorized into six distinct tasks. (1) **General object detection** constitutes the foundation, representing 66.9% of the queries and providing the essential bounding box supervision (83.1%) to help the model achieve precise and dense coordinate alignments. (2) **Grounding user interface** elements (16.5% of queries) enable the model to support embodied agents and graphical user interface navigation tasks. (3) **Natural language referring** comprehension (7.3% of queries) enables the model to link complex linguistic intents to specific spatial regions. (4) **Text localization** (3.6% of queries) ensures that the model can perceive and tightly ground textual information within images. (5) **Document and scene layout grounding** (3.5% of queries) enriches the structural reasoning capabilities of the model. (6) **Point-based localization** tasks (2.2% of queries) further refine the spatial precision of the model for fine-grained predictions.

Table 1: Results on LVIS and COCO. Throughout all tables, "-" means that the information was not reported in the respective papers or the model does not support the corresponding task, **bold** and underline highlight the best and second-best, and BPS (Boxes Per Second) measures decoding throughput.

Method	Throughput	Zero-Shot (LVIS)	LVIS (F1@IoU)			Zero-Shot (COCO)	COCO (F1@IoU)		
			0.5	0.95	Mean		0.5	0.95	Mean
Open-set Specialized Detectors									
Grounding DINO-Swin-T (Liu et al., 2023d)	-	Yes	47.7	22.7	38.8	Yes	69.8	23.0	56.6
Closed-set Specialized Detectors									
Faster RCNN-R50 (Ren et al., 2016)	-	-	-	-	-	No	60.6	7.1	48.1
DETR-R50 (Carion et al., 2020)	-	-	-	-	-	No	65.9	13.6	48.3
Deformable-DETR-R50 (Zhu et al., 2021)	-	-	-	-	-	No	69.7	17.7	54.7
DINO-R50 (Zhang et al., 2022)	-	-	-	-	-	No	68.8	21.1	55.6
DINO-Swin-L (Zhang et al., 2022)	-	-	-	-	-	No	**75.6**	**25.4**	**62.1**
Vision-Language Models									
DeepSeek-VL2-Small (Wu et al., 2024b)	-	-	56.2	21.0	41.8	-	60.9	14.9	45.9
MiMo-VL-7B (Xiaomi Team, 2025)	1.0	-	49.5	8.8	31.4	-	56.5	6.7	35.9
OVIS2.5-2B (Lu et al., 2025b)	1.3	-	54.4	15.8	37.4	-	56.2	10.3	38.7
Qwen3-VL-4B (Bai et al., 2025a)	1.1	-	59.8	20.0	43.5	-	63.0	14.2	46.1
Qwen3-VL-8B (Bai et al., 2025a)	1.0	-	61.5	20.2	44.8	-	62.8	14.0	45.7
Cosmos-Reason2-8B (Cosmos Team, 2025)	1.0	-	56.4	9.8	40.2	-	56.4	9.8	39.3
SEED1.5-VL (Guo et al., 2025)	-	Yes	**65.6**	19.5	46.7	Yes	71.3	14.3	51.4
Rex-Omni-3B (Jiang et al., 2025a)	5.0	Yes	64.3	20.7	46.9	Yes	72.0	15.9	52.9
LocateAnything-3B	**12.7**	Yes	62.3	**31.1**	**50.7**	Yes	70.1	19.3	54.7

4. Experiments

4.1. Training Details and Evaluation Setup

Training Details. We first conduct an initial training on the base VLM with focus entirely on world-knowledge alignment, during which all detection and grounding data are excluded. We then apply a two-stage supervised fine-tuning to the base VLM to train our LocateAnything model. In Stage-1, we incorporate a massive mixture of 138M queries into the overall training data to equip the model with comprehensive grounding and detection capabilities. In Stage-2, we reduce the proportion of general training data to 20% while significantly increasing the proportion of data containing many objects per image (*e.g.*, MOT20Det (Dendorfer et al., 2020), SKU110K (Goldman et al., 2019)) to enhance the model's ability in dense detection. For model ablations, we train all models exclusively on the COCO dataset (Lin et al., 2014) to strictly isolate PBD's architectural benefits from our massive 138M data. Detailed configurations for both the base VLM and the subsequent LocateAnything model training are provided in the **supplementary materials**.

Compared Methods. We compare LocateAnything against three categories of methods. **(1) Specialized detectors**, including representative general detection models such as DETR (Carion et al., 2020) and Deformable-DETR (Zhu et al., 2021), *etc.*, open-set detectors such as Grounding DINO (Liu et al., 2023d), leading document layout analysis model DocLayout-YOLO (Zhao et al., 2024), and text detection model PaddleOCRv5 (Cui et al., 2025). **(2) General-purpose VLMs with grounding capabilities**, including Qwen3-VL (Bai et al., 2025a), DeepSeek-VL2 (Wu et al., 2024b), OVIS2.5 (Lu et al., 2025b), MiMo-VL (Xiaomi Team, 2025), and SEED1.5-VL (Guo et al., 2025), *etc*. These models adopt textual coordinate representations with standard next-token prediction, providing a direct comparison to our parallel box decoding paradigm. **(3) VLM-based detection and grounding specialists**, including Rex-Omni (Jiang et al., 2025a), which is the most related work to ours targeting unified detection and grounding in a VLM framework. For GUI grounding, we also include several domain-specific expert models (Gao et al., 2026; Liu et al., 2025d,e; Xie et al., 2025; Yang et al., 2025b; Ye et al., 2025a; Zhou et al., 2025).

Evaluation Setup. Following the evaluation framework established in Rex-Omni (Jiang et al., 2025a), we conduct a comprehensive assessment across multiple visual perception tasks. Object Detection is evaluated on COCO for common objects, LVIS (Gupta et al., 2019) for long-tailed distributions, and VisDrone (Du et al.,

Table 2: Results on dense object detection benchmark Dense200 and VisDrone.

Method	Score Thresh.	Dense200			VisDrone		
		F1@IoU 0.5	F1@IoU 0.95	F1@IoU Mean	F1@IoU 0.5	F1@IoU 0.95	F1@IoU Mean
Open-set Specialized Detectors							
Grounding DINO-Swin-T (Liu et al., 2023d)	0.25	36.9	**19.7**	33.1	55.2	**3.9**	38.5
Vision-Language Models							
DeepSeek-VL2-Small (Wu et al., 2024b)	-	16.0	3.9	12.7	35.8	1.7	23.3
OVIS2.5-2B (Lu et al., 2025b)	-	17.9	0.0	6.7	21.0	0.1	9.2
MiMo-VL-7B (Xiaomi Team, 2025)	-	29.7	0.4	15.9	27.7	0.3	14.3
Qwen3-VL-4B (Bai et al., 2025a)	-	17.5	2.4	12.5	42.3	1.4	26.0
Qwen3-VL-8B (Bai et al., 2025a)	-	13.5	1.7	9.6	42.8	1.4	25.8
Cosmos-Reason2-8B (Cosmos Team, 2025)	-	25.1	1.1	15.1	40.2	1.3	22.3
SEED1.5-VL (Guo et al., 2025)	-	76.9	5.3	53.2	55.9	0.6	27.4
Rex-Omni-SFT-3B (Jiang et al., 2025a)	-	60.2	10.6	46.4	55.6	1.9	32.4
Rex-Omni-3B (Jiang et al., 2025a)	-	**78.4**	10.3	58.3	61.6	1.5	35.8
LocateAnything-3B	-	74.0	18.5	**58.7**	**63.0**	3.2	**39.9**

Table 3: Results for the GUI Grounding task. The * denotes our reproduced results.

Method	ScreenSpot-Pro												Avg
	Dev.		Creative		CAD		Sci.		Office		OS		
	Text	Icon	Text	Icon	Text	Icon	Text	Icon	Text	Icon	Text	Icon	
InfiGUI-R1-3B (Liu et al., 2025d)	51.3	12.4	44.9	7.0	33.0	14.1	58.3	20.0	65.5	28.3	43.9	12.4	35.7
JEDI-3B (Xie et al., 2025)	61.0	13.8	53.5	8.4	27.4	9.4	54.2	18.2	64.4	32.1	38.3	9.0	36.1
Rex-Omni-3B (Jiang et al., 2025a)	61.7	9.7	52.5	12.6	22.3	9.4	59.0	26.4	63.3	28.3	24.1	15.7	36.8
ScaleCUA-3B (Liu et al., 2025e)	57.8	18.6	38.8	42.9	16.8	32.0	54.3	28.1	47.9	64.6	35.5	**52.0**	40.8
GTA1-7B (Yang et al., 2025b)	62.6	18.2	53.3	17.2	**66.9**	20.7	76.4	31.8	82.5	50.9	48.6	25.9	50.1
Qwen3-VL-30B-A3B* (Bai et al., 2025a)	76.0	24.8	**69.2**	20.3	51.8	15.6	76.4	27.3	80.8	37.7	**75.7**	38.2	53.7
GUI-Owl-7B (Ye et al., 2025a)	76.6	31.0	59.6	27.3	64.5	21.9	79.1	37.3	77.4	39.6	59.8	33.7	54.9
MAI-UI-2B (Zhou et al., 2025)	76.6	32.4	**69.2**	21.7	61.4	23.4	81.2	34.5	**85.9**	39.6	68.2	41.6	57.4
UI-Venus-1.5-2B (Gao et al., 2026)	70.1	43.4	63.6	28.7	54.3	32.8	76.4	38.2	81.9	47.2	73.8	51.7	57.7
GUI-Owl-32B (Ye et al., 2025a)	**84.4**	39.3	65.2	18.2	62.4	28.1	**82.6**	39.1	81.4	39.6	70.1	36.0	58.0
LocateAnything-3B	70.8	**50.3**	60.1	**46.9**	57.9	**40.6**	69.4	**58.2**	77.2	**69.8**	65.4	43.8	**60.3**

2019) and Dense200 (Jiang et al., 2025a) for dense and tiny object scenarios. Language-aware Grounding tasks include Referring Expression Comprehension (REC) on RefCOCOg (Kazemzadeh et al., 2014) and HumanRef (Jiang et al., 2025b). Interactive tasks are evaluated through GUI Grounding on ScreenSpot-Pro (Li et al., 2025b). Additionally, Layout Grounding on DocLayNet (Pfitzmann et al., 2022) and M6Doc (Cheng et al., 2023), along with OCR (text detection and recognition) on TotalText (Ch'Ng and Chan, 2017), are reported together under scene text and document understanding tasks.

The metric for each task is summarized as follows. *(1) Box-based outputs*: For detection, layout, and OCR tasks, a prediction is considered correct (*i.e.*, a true positive) if its Intersection over Union (IoU) with the ground truth exceeds a certain threshold. The F1-score is reported at $IoU = 0.5$, $IoU = 0.95$, and as a mean over thresholds ($mIoU$). *(2) Point-based outputs*: For pointing tasks, a prediction is considered correct if the predicted point falls within the ground-truth segmentation mask or bounding box. We similarly report the F1-score for these point-based outputs based on this correctness criterion.

4.2. Main Results

In this section, we report the accuracy metrics and the throughput (measured in boxes per second, BPS on a single NVIDIA H100 GPU with a batch size of 1) of LocateAnything under the default *Hybrid Mode*. The results of *Fast* and *Slow Mode* are provided in the **supplementary materials**.

High-Quality Multi-Object Detection. Our model exhibits robust generalization in both common and complex dense object detection scenarios. On general detection benchmarks reported in Tab. 1, LocateAnything improves the mean F1 by +3.8% on LVIS and +1.8% on COCO compared to Rex-Omni, despite sharing an identical model size. Crucially, the model effectively learns the generalized spatial distribution, transferring its detection capabilities to unseen, heavily packed object types. This is evidenced by its performance on the dense detection

Table 4: Performance comparison on document layout grounding and OCR tasks.

Method	Score Thresh.	DocLayNet			M6Doc			TotalText		
		F1@IoU 0.5	F1@IoU 0.95	F1@IoU Mean	F1@IoU 0.5	F1@IoU 0.95	F1@IoU Mean	F1@IoU 0.5	F1@IoU 0.95	F1@IoU Mean
Specialized Detectors										
DocLayout-YOLO (Zhao et al., 2024)	0.3	**91.2**	**52.1**	**81.1**	-	-	-	-	-	-
PaddleOCRv5 (Cui et al., 2025)	-	-	-	-	-	-	-	40.2	0.7	25.7
Vision-Language Models										
SEED1.5-VL (Guo et al., 2025)	-	54.9	4.3	28.7	48.0	3.4	28.0	35.0	0.3	19.5
Qwen3-VL-4B (Bai et al., 2025a)	-	60.8	8.2	37.2	30.6	4.9	19.0	55.4	3.6	36.1
Qwen3-VL-8B (Bai et al., 2025a)	-	54.7	6.7	34.1	37.2	4.9	22.7	**59.4**	2.7	37.3
Rex-Omni-3B (Jiang et al., 2025a)	-	89.5	28.4	70.7	76.3	18.7	55.6	56.6	3.9	40.6
LocateAnything-3B	-	91.1	35.8	76.8	**90.6**	**25.8**	**70.1**	58.9	**5.1**	**43.3**

Table 5: Evaluation results on referring expression comprehension benchmarks.

Method	Score Thresh.	HumanRef			RefCOCOg val			RefCOCOg test		
		F1@IoU 0.5	F1@IoU 0.95	F1@IoU Mean	F1@IoU 0.5	F1@IoU 0.95	F1@IoU Mean	F1@IoU 0.5	F1@IoU 0.95	F1@IoU Mean
Open-set Specialized Detector										
Grounding DINO-Swin-T (Liu et al., 2023d)	0.25	28.0	16.5	25.2	52.9	20.9	45.9	53.8	22.9	46.8
Vision-Language Model										
DeepSeek-VL2-Tiny (Wu et al., 2024b)	-	39.1	16.9	31.4	67.4	16.1	50.5	69.3	16.9	52.1
OVIS2.5-2B (Lu et al., 2025b)	-	70.6	12.3	50.0	87.4	29.3	73.4	87.6	30.5	73.8
MiMo-VL-7B (Xiaomi Team, 2025)	-	77.6	26.4	63.4	84.9	14.4	65.3	84.6	14.9	65.5
DeepSeek-VL2-Small (Wu et al., 2024b)	-	72.0	46.5	64.7	**92.4**	**45.6**	**81.4**	**91.8**	**47.0**	**81.6**
Qwen3-VL-4B (Bai et al., 2025a)	-	77.7	54.9	71.1	88.0	34.0	74.7	87.6	33.9	74.6
Qwen3-VL-8B (Bai et al., 2025a)	-	78.6	55.7	72.0	88.6	33.4	74.9	88.6	33.8	75.2
SEED1.5-VL (Guo et al., 2025)	-	**88.2**	60.0	**81.6**	84.7	30.9	71.9	85.2	32.1	73.2
Rex-Omni-3B (Jiang et al., 2025a)	-	85.4	65.4	79.9	86.6	35.3	73.6	86.8	36.6	74.3
LocateAnything-3B	-	82.9	**68.8**	78.7	88.6	41.5	76.7	88.8	43.4	77.6

benchmarks in Tab. 2, where it achieves 39.9 mean F1 on VisDrone, substantially outperforming Rex-Omni which scores 35.8. Similarly, it reaches a competitive 58.7 mean F1 on Dense200, demonstrating superior boundary delineation and instance separation in heavily overlapping environments.

Precise Open-World Localization Ability. LocateAnything demonstrates exceptional fine-grained localization capabilities across diverse open-world benchmarks, including user interface grounding, document layout parsing, and referring expression comprehension. As shown in Tab. 3, on the ScreenSpot-Pro (Li et al., 2025b), it achieves a SOTA mean F1 of 60.3, surpassing generalist VLMs like Qwen3-VL-30B-A3B and specialized models tailored for UI tasks such as GUI-Owl-32B. Furthermore, in document understanding tasks detailed in Tab. 4, LocateAnything establishes a new standard by reaching 76.8 and 70.1 mean F1 on DocLayNet and M6Doc, respectively, outperforming Rex-Omni by substantial margins. This precise spatial reasoning extends to complex referring tasks, as shown in Tab. 5, where the model seamlessly aligns nuanced human intents with visual regions, achieving 78.7 mean F1 on the HumanRef benchmark and remaining highly competitive on RefCOCOg against top-tier models.

Superior Decoding Speed. A key advantage of our model is its drastically reduced decoding steps. As shown in Tab. 1, our model achieves 12.7 BPS under the default hybrid mode, over 10× faster than textual-based Qwen3-VL (1.1 BPS) and 2.5× faster than quantized-based Rex-Omni (5.0 BPS).

4.3. Ablation Study

We conduct ablation studies on the COCO dataset to validate our core designs. The results are shown in Tab. 6 and Fig. 7.

Coordinate Representation. As Tab. 6(a) shows, under the NTP paradigm, Textual and Quantized representations yield sub-optimal performance (49.1 and 50.1 mean F1, respectively) due to forced token-by-token generation. Our PBD (Slow Mode) achieves the highest F1-score of 52.1, proving that a box-aligned formulation provides stronger supervision for spatial reasoning than 1D serialization, without sacrificing throughput.

Table 6: Ablation Studies on the COCO dataset. We decouple the analysis into three aspects: (a) coordinate representation, (b) block-based MTP Formulation, and (c) effectiveness of our on-demand decoding modes and loss design. Throughput is measured in boxes per second. For brevity, we report the Average metric across IoU thresholds for Recall (R), Precision (P), and F1 Score. "B" indicates block size in MTP.

(a) Coordinate Representations

Method	Throughput	R	P	F1
Textual	1.3	45.7	52.3	49.1
Quantized	3.9	48.2	52.2	50.1
PBD (Slow)	3.9	**49.4**	**55.2**	**52.1**
PBD (Fast)	**16.9**	45.6	54.6	49.6
PBD (Hybrid)	13.2	48.7	54.8	51.6

(b) MTP Formulations

Method	Throughput	R	P	F1
SDLM-B4 (Liu et al., 2025c)	5.2	**45.4**	48.1	46.5
SDLM-B6 (Liu et al., 2025c)	5.5	45.1	47.5	46.1
SDLM-B8 (Liu et al., 2025c)	6.7	44.7	47.2	45.8
Block Diff-B6 (Arriola et al., 2025)	4.7	45.1	44.3	44.8
PBD (Fast)	**16.9**	45.2	**54.6**	**49.6**

(c) Decoding Modes & Losses

$\mathcal{L}_{\text{ntp}}$	$\mathcal{L}_{\text{blk}}$	Mode	Throughput	R	P	F1
✓		Slow	3.9	48.2	52.2	50.1
	✓	Fast	16.7	45.6	49.0	47.2
✓	✓	Slow	3.9	**49.4**	**55.2**	**52.1**
✓	✓	Fast	**16.9**	45.6	54.6	49.6
✓	✓	Hybrid	13.2	48.7	54.8	51.6

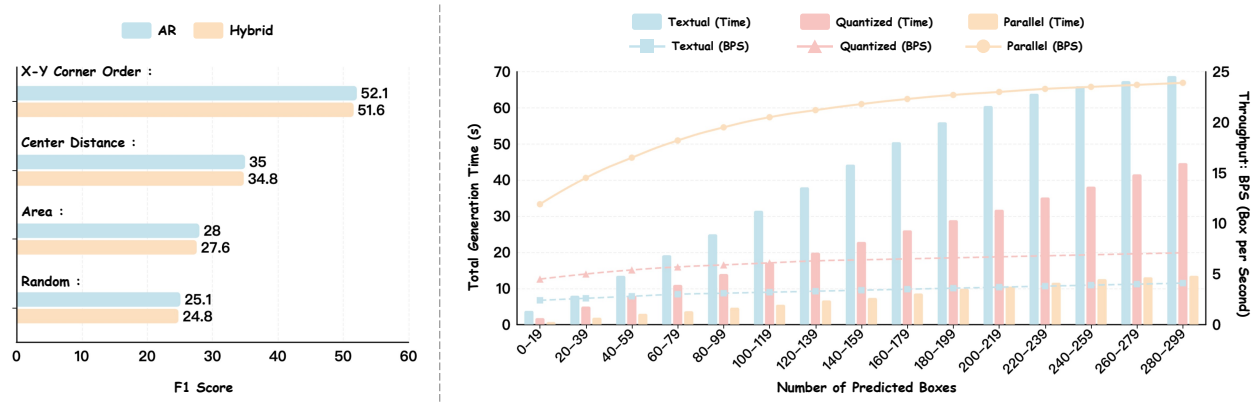


Figure 7: **Ablation Study on Box Ordering and Decoding Speed. Left:** Effect of different box sorting strategies on the F1-score. **Right:** Comparison of Generation Time (bars) and Throughput (lines) across varying numbers of predicted boxes for Textual, Quantized, and Parallel box decoding.

MTP Formulation. Tab. 6(b) compares our box-aligned MTP against existing structure-agnostic MTP formulations. Methods like SDLM and Block Diffusion force the model to learn spurious, unaligned cross-boundary patterns, suffering from lower accuracy and limited acceleration (*e.g.*, SDLM-B6 achieves 46.1 F1-score at 5.5 BPS). Furthermore, structure-agnostic methods (*e.g.*, SDLM-B4, B6, B8) exhibit a strict speed-accuracy trade-off, where increasing the block size yields only marginal throughput gains while consistently degrading the F1-score. In contrast, our PBD strictly aligns MTP blocks with structured bounding box units, dramatically outpacing existing methods in throughput (16.9 BPS) while improving the mean F1 to 49.6.

Decoding Mode. Tab. 6(c) ablates the impact of our dual-formulation training ($\mathcal{L}_{\text{ntp}}$ and $\mathcal{L}_{\text{blk}}$). Training with isolated losses limits the model's potential; joint training successfully pushes the Slow Mode upper bound from 50.1 to 52.1 F1-score. During inference, Fast Mode (MTP) maximizes throughput (16.9 BPS) but induces accuracy drops in complex scenes. Hybrid Mode seamlessly resolves this trade-off, preserving most speed gains (13.2 BPS) while achieving robust, high-precision localization (51.6 F1-score).

Box Output Order. We investigate four spatial sorting strategies in Fig. 7 (left): *X-Y Corner Order* (sorting by the x-coordinate of the left-top corner, then by the y-coordinate), *Center Distance* (the distance of the bounding box center point to the origin), *Area* (sorted from largest to smallest), and *Random* (shuffled randomly). Results show X-Y Corner Order yields the highest F1-score. We take this setting as default in dataset construction.

Throughput. We compare generation time and throughput with NTP methods in Fig. 7 (right). As target boxes increase from 20 to 300, NTP methods suffer from a severe latency bottleneck. In contrast, the Parallel method exhibits little increase in generation time, increasing throughput from 12 BPS to ∼25 BPS in dense scenes. These findings confirm that PBD effectively breaks the decoding bottleneck, achieving a $2\times$ to $6\times$ speedup.

4.4. Qualitative Results

Fig. 8 visualizes representative grounding results of our model. Visual comparisons with other methods are provided in the **supplementary materials**. We observe three consistent behaviors. **(i) Compositional**

Figure 8: **Qualitative results**. Each row shows test cases with varying numbers of target objects and diverse box scales. Different colors indicate different query categories, including attribute, part, reasoning, and spatial queries. Our model consistently localizes targets across diverse scene domains, arbitrary image resolutions, free-form textual queries, and an arbitrary number of objects, demonstrating strong robustness.

grounding: our model handles attribute/part/spatial/reasoning-style queries well with consistent spatial alignment, supported by the diversity and coverage of our training data. **(ii) Robustness to large instance counts:** as targets grow from sparse to crowded settings, the predicted boxes remain structured and accurate, reflecting the precision of our box-level decoding. This robustness is further strengthened by our Stage-2 training that emphasizes many-object images, improving dense localization in practice. Moreover, our Hybrid Mode maintains most of the parallel decoding speed while improving output stability in multi-instance generation. **(iii) Reliable localization in clutter:** boxes stay compact and well-separated under occlusion, repetitive textures, and grid-like dense layouts. Our hybrid inference mode further stabilizes these hard cases by detecting unreliable parallel blocks and falling back to NTP re-decoding when needed.

5. Conclusion

We presented LocateAnything, a unified framework that reformulates visual grounding and detection in VLMs via *Parallel Box Decoding*. By elevating geometric elements to atomic units rather than 1D streams, LocateAnything aligned the training supervision with the inherently coupled nature of spatial coordinates. With massive 138M text-image training queries and a flexible on-demand inference mechanism, LocateAnything not only delivered SOTA accuracy across diverse tasks, but also achieved up to a $2.5\times$ speedup over competitive methods. Our method provided a practical and scalable route for real-time visual perception, opening the door to deploying general-purpose VLMs in latency-sensitive embodied robotics and interactive agents.

Limitation. Currently, our model is primarily trained with supervised fine-tuning. Reinforcement learning is an important next step to further optimize the block-level decoding policy, reduce fallback frequency, and encourage effective exploration in hard dense/long-tail cases, which could improve both robustness and worst-case decoding speed. We leave it for future work.

Acknowledgement. The authors would like to thank the valuable discussions and input from Qing Jiang, Amala Sanjay Deshmukh, Karan Sapra, Mingjie Liu, Yi Dong, Pavlo Molchanov, Yonggan Fu, Collin McCarthy,

Mike Ranzinger, Greg Heinrich, Wonmin Byeon, Yexuan Li, Chi-Pin Huang, Fu-En Yang, Frank Wang, Jin Huang, Le An, Jaehun Jung, Shaokun Zhang, Hao Zhang, Johan Bjoerck, Jim Fan, Patrick Langechuan Liu, Sifei Liu, Xiaolong Li, Paris Zhang, Yilin Zhao, Subhashree Radhakrishnan, Shiyi Lan, Jose Alvarez, Sanja Fidler, Yan Wang, Xiaodong Yang, Yin Cui, Tsung-Yi Lin, Padmavathy Subramanian and more. We would also like to thank the NVIDIA infra, legal and data teams, including Xinyou Ma, Katherine Cheung, Timo Roman, and Yao Xu for their prompt and helpful support. Finally, the authors would like to additionally acknowledge the following teams, including Nemotron-Diffusion, Nemotron VLM, Cosmos, GR00T, Alpamayo, Gigas and Metropolis, for the engagement and downstream applications.

A. Training and Inference Configurations

A.1. Training Details

In this section, we provide extended details regarding our data mixture strategy, the multiphase training pipeline of our base VLM and the LocateAnything model. We also elaborate on two key system-level techniques that are critical for efficient training under our dual-formulation design: *Stream Packing* for maximizing GPU utilization, and *MagiAttention* (Zewei and Yunpeng, 2025) for natively supporting the heterogeneous attention masks required by our NTP+MTP joint training.

To provide a comprehensive overview of our entire training pipeline, Tab. 7 summarizes the detailed optimization hyperparameters and configurations across all four progressive stages of **LocateAnything**.

Table 7: Detailed configuration for each training stage of **LocateAnything**.

Stages	**Stage 1**	**Stage 2**	**Stage 3**	**Stage 4**
Objective	**World Knowledge Injection**		**Detection & Grounding Enhancement**	
Dataset	Caption	General VQA	Detection & Grounding	20% Previous + Dense
Learning Rate	2×10^{-4}	4×10^{-5}	4×10^{-5}	1×10^{-5}
Optimizer	AdamW	AdamW	AdamW	AdamW
Weight Decay	0.01	0.01	0.01	0.01
LR Schedule	Cosine	Cosine	Cosine	Cosine
Max Sequence Length	32768	32768	25600	25600
Trainable Components	MLP	All	All	All
Number of GPUs	64	256	256	256
Training Steps	2000	20000	25000	5000

A.1.1. Base VLM Training (World Knowledge Injection)

To establish a robust foundational understanding of world knowledge before introducing specialized detection and grounding tasks, we first pretrain our base VLM. This initial alignment phase strictly excludes any detection or bounding-box grounding data and is divided into the first two progressive stages. (1) **Stage 1 (Visual Concept Initialization):** In this stage, the model is trained exclusively on caption-related datasets, detailed in the "Captioning & Knowledge" category of Tab. 8. This enables the native any-resolution visual encoder to align fundamental visual features with textual descriptions effectively. (2) **Stage 2 (Comprehensive Multimodal Learning):** Building upon the basic captioning capability, we expand the training corpus to encompass all datasets listed in Tab. 8. This comprehensive mixture spans a wide spectrum of domains, including Mathematics & Code, Science, Chart & Table reasoning, extensive OCR tasks (Naive OCR and OCR QA), General VQA, Text-only instruction tuning, and basic Counting. Fully integrating these diverse datasets ensures the base model develops strong reasoning and comprehensive multimodal capabilities.

A.1.2. LocateAnything Fine-Tuning (Detection and Grounding Enhancement)

Following the initial world-knowledge alignment, we then train the LocateAnything model using a carefully designed two-stage SFT strategy tailored for fine-grained detection and grounding. This constitutes the final two stages of our pipeline (leveraging the data presented in Fig. 5 of the main text). (1) **Stage 3 (Comprehensive Detection and Grounding):** We incorporate a massive mixture of 138M queries into the overall training data to equip the model with comprehensive grounding and detection capabilities. During this stage, all model components are fully unfrozen and trained. We set the maximum sequence length to 25,600 and employ a learning rate of 4×10^{-5} with a Cosine schedule. (2) **Stage 4 (Dense Detection Enhancement):** To further boost the model's recall in dense scenes, we reduce the proportion of general training data to 20% while significantly increasing the proportion of data containing many objects per image (*e.g.*, MOT20Det, SKU110K).

Table 8: **Datasets used for the initial world-knowledge alignment.** We pretrain the base VLM on this diverse mixture of datasets across various domains to ensure broad coverage of general knowledge. Specifically, Stage-1 incorporates only the caption-related datasets shown here. In Stage-2, all datasets listed in this table are fully integrated into the training process to build robust, comprehensive multimodal capabilities.

Category	Dataset
Captioning & Knowledge	ShareGPT4o OpenGVLab (2024), KVQA Shah et al. (2019), Movie-Posters skvarre (2024), Google-Landmark Weyand et al. (2020), WikiArt HugGAN (2024), Weather-QA Ma et al. (2024a), Coco-Colors hazal karakus (2024), music-sheet EmileEsmaili (2024), SPARK Yu et al. (2024b), Image-Textualization Pi et al. (2024), SAM-Caption PixArt-alpha (2024), Tmdb-Celeb-10k Ashraq (2024), CC3M Sharma et al. (2018), pixmo-cap Deitke et al. (2025), Multi-UI Liu et al. (2024b), RICO Deka et al. (2017)
Mathematics & Code	GeoQA+ Cao and Xiao (2022), MathQA Yu et al. (2023a), CLEVR-Math/Super Li et al. (2023b); Lindström and Abraham (2022), Geometry3K Lu et al. (2021a), MAVIS-math-rule-geo Zhang et al. (2024d), MAVIS-math-metagen Zhang et al. (2024d), InterGPS Lu et al. (2021b), Raven Zhang et al. (2019a), GEOS Seo et al. (2015), UniGeo Chen et al. (2022a), Design2Code Si et al. (2025), OpenMathInstruct Toshniwal et al. (2024)
Science	AI2D Kembhavi et al. (2016), ScienceQA Lu et al. (2022a), TQA Kembhavi et al. (2017), PathVQA He et al. (2020), SciQA Auer et al. (2023), Textbooks-QA, VQA-RAD Lau et al. (2018), VisualWebInstruct TIGER-Lab (2024), PMC-VQA Zhang et al. (2023a)
Chart & Table	ChartQA Masry et al. (2022), MMC-Inst Liu et al. (2023b), DVQA Kafle et al. (2018), PlotQA Methani et al. (2020), LRV-Instruction Liu et al. (2023a), TabMWP Lu et al. (2022b), UniChart Masry et al. (2023), Vistext Tang et al. (2023), TAT-DQA Zhu et al. (2022), VQAonBD VQAonDB, FigureQA Kahou et al. (2017), Chart2Text Kantharaj et al. (2022), RobuT-{Wikisql, SQA, WTQ} Zhao et al. (2023), MultiHiertt Zhao et al. (2022), MMTab Zheng et al. (2024a)
Naive OCR	SynthDoG Kim et al. (2022), MTWI He et al. (2018), LVST Sun et al. (2019), SROIE Huang et al. (2019), FUNSD Jaume et al. (2019), Latex-Formula OleehyO (2024), IAM Marti and Bunke (2002), Handwriting-Latex aida-pearson (2023), ArT Chng et al. (2019), CTW Yuan et al. (2019), ReCTs Zhang et al. (2019b), COCO-Text Veit et al. (2016), SVRD Yu et al. (2023b), Hiertext Long et al. (2023), RoadText Tom et al. (2023), MapText Li et al. (2024b), CAPTCHA parasam (2024), Est-VQA Wang et al. (2020), HME-100K TAL (2023), TAL-OCR-ENG TAL (2023), TAL-HW-MATH TAL (2023), IMGUR5K Krishnan et al. (2023), ORAND-CAR Diem et al. (2014), Invoices-and-Receipts-OCR mychen76 (2024), Chrome-Writting Mouchère et al. (2016), IIIT5k Mishra et al. (2012), K12-Printing TAL (2023), Memotion Ramamoorthy et al. (2022), Arxiv2Markdown, Handwritten-Mathematical-Expression Azu (2023), WordArt Xie et al. (2022), RenderedText wendlerc (2024), Handwriting-Forms ift (2024)
OCR QA	DocVQA Clark and Gardner (2018), InfoVQA Mathew et al. (2022), TextVQA Singh et al. (2019), ArxivQA Li et al. (2024a), ScreencQA Hsiao et al. (2022), DocReason mPLUG (2024), Ureader Ye et al. (2023), FinanceQA Sujet AI et al. (2024), DocMatrix Laurençon et al. (2024a), A-OKVQA Schwenk et al. (2022), Diagram-Image-To-Text Kamizuru00 (2024), MapQA Chang et al. (2022), OCRVQA Mishra et al. (2019), ST-VQA Biten et al. (2019), SlideVQA Tanaka et al. (2023), PDF-VQA Ding et al. (2023), SQuAD-VQA, VQA-CD Mahamoud et al. (2024), Block-Diagram shreyanshu09 (2024), MTVQA Tang et al. (2024), ColPali Faysse et al. (2024), BenthamQA Mathew et al. (2021), VSR Zhang et al. (2021), pixmo-docs Deitke et al. (2025)
General VQA	LLaVA-150K Liu et al. (2023c), LVIS-Instruct4V Wang et al. (2023a), ALLaVA Chen et al. (2024), Laion-GPT4V LAION (2023), LLAVAR Zhang et al. (2023b), SketchyVQA Tu et al. (2023), VizWiz Gurari et al. (2018), IDK Cha et al. (2024), AlfworldGPT, LNQA Pont-Tuset et al. (2020), Face-Emotion FastJobs (2024), SpatialSense Yang et al. (2019), Indoor-QA keremberke (2024), Places365 Zhou et al. (2017), MMinstruct Liu et al. (2024c), DriveLM Sima et al. (2023), YesBut Nandy et al. (2024), WildVision Lu et al. (2024), LLaVA-Critic-113k Xiong et al. (2024), RLAIF-V Yu et al. (2024a), VQAv2 Goyal et al. (2017), MMRA Wu et al. (2024a), KONIQ Hosu et al. (2020), MMDU Liu et al. (2024d), Spot-The-Diff Jhamtani and Berg-Kirkpatrick (2018), Hateful-Memes Kiela et al. (2020), COCO-QA Ren et al. (2015), NLVR Suhr et al. (2017), Mimic-CGD Laurençon et al. (2024b), Datikz Belouadi et al. (2023), Chinese-Meme Contributors (2024), IconQA Lu et al. (2021c), Websight Laurençon et al. (2024c), OmniAlign Zhao et al. (2025), pixmo-cap-qa Deitke et al. (2025), pixmo-ask-model-anything Deitke et al. (2025), Cauldron Laurençon et al. (2024b)
Text-only	Orca Lian et al. (2023), Orca-Math Mitra et al. (2024), OpenCodeInterpreter Zheng et al. (2024b) MathInstruct Yue et al. (2023), WizardLM Xu et al. (2023), TheoremQA Chen et al. (2023b), OpenHermes2.5 Teknium (2023), NuminaMath-CoT LI et al. (2024), Python-Code-25k flytech (2024), Infinity-Instruct BAAI (2024), Python-Code-Instructions-18k-Alpaca iamtarun (2024), Ruozhiba LooksJuicy (2024), InfinityMATH Zhang et al. (2024a), StepDPO Lai et al. (2024b), TableLLM Zhang et al. (2024e), UltraInteract-sft Yuan et al. (2024)
Counting	FSC147 Ranjan et al. (2021), TallyQA Acharya et al. (2019)

All components remain trainable, and the maximum sequence length is maintained at 25,600. The learning rate is decayed to 1×10^{-5}.

A.1.3. Stream Packing

A key challenge for training with our dual-formulation (NTP + MTP) design is that different samples, after block-wise expansion, exhibit highly variable sequence lengths. Naïve padding-based batching leads to significant GPU memory waste and low arithmetic utilization. To address this, we adopt an *online stream packing* strategy that dynamically assembles multiple variable-length samples into a single, densely packed sequence of a target budget (*e.g.*, 36,864 tokens). Concretely, our packing pipeline operates through three core mechanisms. First, via **Weighted Sampling**, an infinite iterator draws samples from multiple heterogeneous datasets according to pre-specified mixing weights. Second, utilizing **Best-Fit Buffering**, a fixed-size buffer (default size 32) stores pending samples. When assembling a batch, the packer first scans the buffer for the *largest* sample that still fits into the remaining token budget—a best-fit decreasing heuristic that empirically yields $>$95% packing efficiency. If no buffered sample fits, a freshly drawn sample is either appended directly (if it fits) or placed into the buffer for future use. Third, through **Big-Rocks-First Seeding**, after yielding a completed batch, the packer seeds the next batch with the *largest* sample currently in the buffer, ensuring that oversized samples are never starved. Each packed sequence carries a `sub_sample_lengths` tensor that records the constituent sample boundaries. This metadata is consumed downstream by the attention kernel to construct the correct per-sample attention mask within the packed sequence, preventing cross-contamination between unrelated samples.

A.1.4. MagiAttention for Heterogeneous Mask Training

Our dual-formulation training produces a *heterogeneous* attention mask that combines standard causal attention (for the NTP stream) with block-causal and bidirectional intra-block patterns (for the MTP stream), all within a single packed sequence potentially containing multiple samples. When further combined with stream packing, the resulting attention mask becomes highly irregular and sample-dependent, making it incompatible with conventional Flash-Attention kernels that assume a uniform causal or full-attention pattern.

To efficiently handle this, we leverage **MagiAttention** (Zewei and Yunpeng, 2025), a distributed attention framework designed for ultra-long contexts with heterogeneous masks. Together, stream packing and MagiAttention form a synergistic training infrastructure: packing maximizes token-level utilization within each GPU, while MagiAttention ensures that the resulting heterogeneous attention masks are handled both correctly and efficiently across the distributed training cluster.

A.2. Inference Details

We provide a detailed description of the inference pipeline, including the generation modes, the semi-autoregressive generation loop, the box-aware decoding strategies, and the hyperparameter configurations used across all evaluations.

A.2.1. Generation Hyperparameters

We employ nucleus sampling with a temperature of 0.7 and top-p of 0.9 to balance diversity and precision. A repetition penalty of 1.1 is applied to discourage duplicate predictions. KV cache is enabled throughout inference to avoid redundant computation. The block size for MTP generation is set to 6 (*i.e.*, $n_{\text{future}} = 6$), meaning each parallel decoding step predicts up to 6 tokens simultaneously. The maximum number of newly generated tokens is set to $8{,}192$. All models are evaluated in BF16 precision with a batch size of 1.

A.2.2. KV Cache Management

After each MTP step, the KV cache is truncated to include only the positions corresponding to actually committed tokens (*i.e.*, the prefix up to the current generation frontier). The mask tokens and the duplicated anchor token are evicted, ensuring that subsequent steps attend only to the ground-truth generation history. This truncation is essential for maintaining consistency between the causal prefix seen during training and the KV cache state during inference.

Table 9: Overview of supported perception tasks and their corresponding prompt templates. `[PHRASE]` denotes a free-form natural language description, and `[CATEGORIES]` denotes a comma-separated list of category names.

Task	Output	Question Template
Object Detection	Box	Locate all the instances that matches the following description: `[CATEGORIES]`.
Phrase Grounding	Single Box	Locate a single instance that matches the following description: `[PHRASE]`.
	Multiple Boxes	Locate all the instances that match the following description: `[PHRASE]`.
Text Grounding	Box	Please locate the text referred as `[PHRASE]`.
Scene Text Detection	Box	Detect all the text in box format.
Document Layout Analysis	Box	Detect all the objects in the image that belong to the category set: `[CATEGORIES]`.
GUI Grounding	Box	Locate the region that matches the following description: `[PHRASE]`.
	Point	Point to: `[PHRASE]`.
Pointing	Point	Point to: `[PHRASE]`.

B. LocateAnything-Data Construction

B.1. Leveraging Existing Open-Source Data

We begin by collecting high-quality detection and grounding datasets from the open-source community and performing unified format cleaning and normalization. As illustrated in Fig. 6 of the main paper, the collected data span six domains, covering diverse visual scenarios.

Except for GroundCUA (Feizi et al., 2025a), we use the original labels for all other GUI datasets. The GroundCUA dataset, however, requires additional processing because its original labels typically correspond only to short descriptions of UI elements. To enrich the grounding queries for this specific dataset, we augment the GroundCUA annotations using Qwen3-VL (Bai et al., 2025a). Specifically, given the original *bbox*, *label*, and *category*, we first render the target bounding box on the screenshot and crop a local region around it. The full screenshot, the cropped region, together with the label, category, and platform metadata are then provided to Qwen3-VL (Bai et al., 2025a). After determining whether the target element is visually identifiable, the model generates natural language descriptions from three complementary perspectives: *appearance*, describing visual attributes such as color, shape, iconography, or textual content; *spatial*, describing the element's relative position with respect to other UI components; and *functional*, describing the user intent or interaction semantics associated with the element. Through this process, the original discrete text labels of GroundCUA are transformed into richer, multi-dimensional grounding queries that are both descriptive and interpretable.

Referring and grounding datasets themselves are relatively limited in scale. To address this, we aggregate several widely used benchmarks, including Flickr30k Entities (Plummer et al., 2016), gRefCOCO (He et al., 2023), RefCOCO (Yu et al., 2016), HumanPart (Yu et al., 2016), and HumanRef (Jiang et al., 2025a). In addition, we incorporate large-scale detection datasets such as OpenImages (Kuznetsova et al., 2020), Objects365 (Shao et al., 2019), and images collected from Unsplash. These datasets serve as raw sources for constructing our multi-target grounding data engine, as discussed in Sec. B.2.

Another critical issue in existing detection and grounding datasets is that they almost exclusively contain *positive* samples. Training on such data can lead to hallucination behaviors, where the model predicts bounding boxes even when the query is unrelated to the image. To mitigate this issue, we explicitly construct negative samples across domains. The proportion of negative queries varies depending on the domain statistics (see

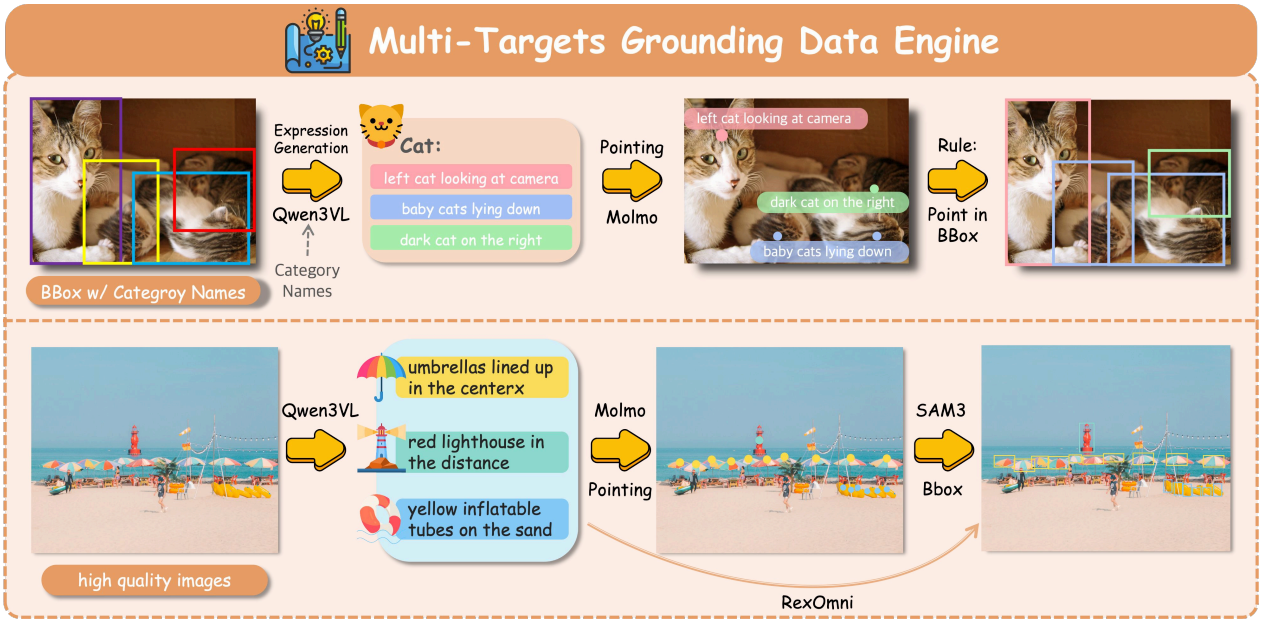


Figure 9: **Data engine for multi-targets grounding. Top:** For detection datasets with gt boxes, we use each box category as a prompt to Qwen3-VL (Bai et al., 2025a) to synthesize detailed object-centric queries, including attributes, spatial relations, and reasoning cues. These queries are then fed to Molmo (Deitke et al., 2025) to predict candidate points, from which we retain the points falling inside the corresponding gt boxes as reliable supervision. **Bottom:** For a large collection of high-quality unlabeled images, Qwen3-VL directly generates diverse queries from the image. Such queries can be used to prompt Molmo for point prediction, followed by SAM 3 (Carion et al., 2025) to produce boxes, or directly prompt Rex-Omni (Jiang et al., 2025a) to generate boxes. All generated boxes are finally post-verified by Qwen3-VL.

Tab. 10). Concretely, we generate queries referring to objects that do not exist in the image, and assign them the *negative block* described in Fig. 3 of the main paper. This design enables the model to learn to abstain when no valid grounding target is present.

B.2. Multi-Targets Grounding Data Engine

Existing open-source grounding datasets are relatively limited in scale and diversity. To construct a large-scale multi-target grounding dataset, we design a data engine that automatically synthesizes grounding annotations from both labeled detection data and unlabeled images, as illustrated in Fig. 9.

From Detection Datasets:
We first leverage high-quality detection datasets such as Open Images (Kuznetsova et al., 2020) and Objects365 (Shao et al., 2019). For each ground-truth bounding box, we use its category label as a prompt to Qwen3-VL (Bai et al., 2025a) to generate a set of detailed object-centric queries, including attributes, spatial relations, and reasoning cues. These queries are then used to prompt Molmo (Deitke et al., 2025) to predict candidate points. Since the ground-truth boxes are known, we retain only the points that fall inside the corresponding bounding boxes, which serve as reliable grounding supervision.

From Unlabeled Images:
To further expand the diversity of grounding targets, we additionally collect large amounts of high-quality unlabeled images from Unsplash and SA-1B (Kirillov et al., 2023). For each image, Qwen3-VL directly generates a diverse set of natural language queries describing potential objects or regions. These queries can be used to prompt Molmo (Deitke et al., 2025) to predict points, which are subsequently converted into bounding boxes using SAM 3 (Carion et al., 2025). Alternatively, the queries can directly prompt Rex-Omni (Jiang et al., 2025a) to predict bounding boxes.

Table 10: Statistics of the collected data across six domains. We report the total number of queries and negative samples, together with the maximum and mean numbers of targets and categories per query (/ Q), and targets per image (/ I). *Query length* measures the number of words in the target description after removing template text, reflecting the actual linguistic content used to describe grounding targets.

Domain	#Queries	#Negative	Targets / Q		Categories / Q		Query Length		Targets / I	
			Max	Mean	Max	Mean	Max	Mean	Max	Mean
Detection	93,351,373	21,021,509	745	6.29	43	2.47	251	24.19	3,725	30.68
GUI	23,009,535	0	14	1.03	14	1.03	351	4.07	8,690	7.95
Referring	10,141,597	93,396	818	2.12	1	0.89	53	5.48	6,938	9.65
OCR	5,052,040	0	2,337	11.89	1,258	10.4	51	1.17	2,337	28.67
Layout	4,859,914	1,384,804	176	4.92	15	1.31	30	2.2	880	21.17
Pointing	3,148,098	353,366	675	3.25	1	0.89	189	2.63	1,575	14.92

To ensure annotation quality, all generated boxes are finally verified by Qwen3-VL(Bai et al., 2025a) through a post-checking stage, filtering out inconsistent predictions.

B.3. Task-Specific Prompt Design

As detailed in Tab. 9, we present a comprehensive overview of the versatile perception tasks supported by our unified framework, alongside their corresponding output formats and question templates. To seamlessly integrate diverse visual grounding and detection capabilities, we design specific textual prompts for each task. The model handles a wide spectrum of region-based tasks that output bounding boxes, including Object Detection, Text Grounding, Scene Text Detection, and Document Layout Analysis. Furthermore, it supports fine-grained localization tasks such as Pointing, which outputs specific coordinate points. For complex referring and interactive tasks like Phrase Grounding and GUI Grounding, the framework flexibly predicts either single/multiple boxes or points depending on the user's intent. Within the prompt templates, `[PHRASE]` represents a free-form natural language description, while `[CATEGORIES]` denotes a comma-separated list of target category names. This unified prompting strategy enables the model to effectively bridge natural language instructions with precise spatial coordinate decoding.

B.4. Data Statistics and Distribution

We analyze the statistical characteristics of the collected dataset. Tab. 10 summarizes the dataset statistics across six domains. In total, the dataset contains over 139M queries with more than 22M negative samples.

Our dataset also exhibits strong multi-target grounding characteristics. The number of targets associated with each query varies substantially across domains. As illustrated in Fig. 10, the distribution of targets per query follows a long-tailed pattern: most queries correspond to a small number of targets, while a non-negligible portion involve a large number of instances.

We further analyze the linguistic properties of the queries. As shown in Fig. 11, query length varies across domains, reflecting different grounding paradigms and language patterns used to describe visual targets.

Overall, these statistics highlight the scale, diversity, and multi-target nature of our dataset, which together provide a strong foundation for training models capable of handling heterogeneous visual domains and complex language queries.

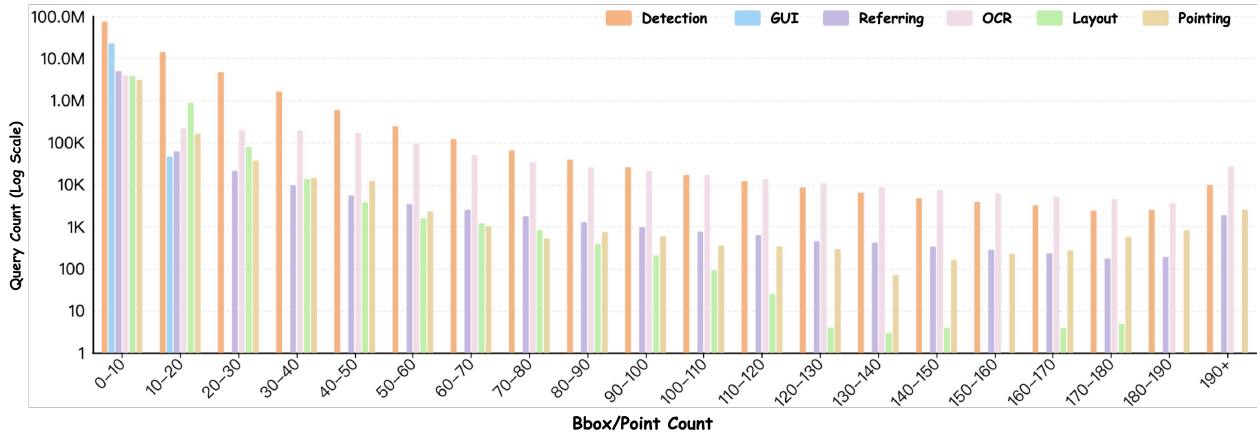


Figure 10: Distribution of the number of targets per query across different domains. The x-axis shows the number of targets associated with a query, while the y-axis (log scale) indicates the number of queries.

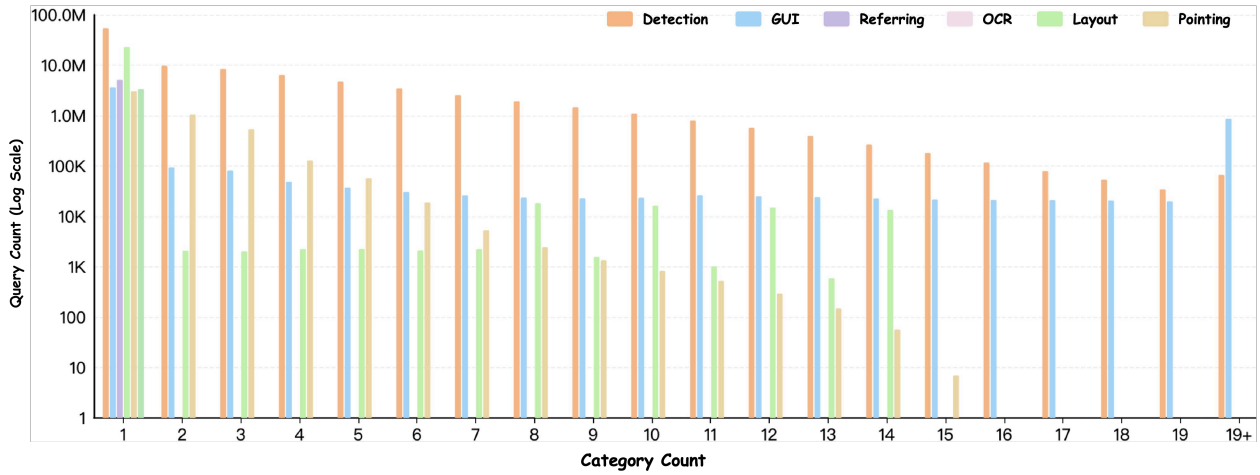


Figure 11: Distribution of query length across domains. The x-axis represents the number of words in the target description (excluding template text), and the y-axis (log scale) shows the number of queries.

C. Additional Experiments

C.1. Results on Pointing Tasks

To further evaluate the fine-grained spatial perception of our model, we benchmark LocateAnything-3B on point-based localization tasks, where the model must predict a point that falls within the target's bounding box or segmentation mask. As detailed in Tab. 11, LocateAnything-3B (evaluated under Hybrid Mode) achieves state-of-the-art results across a diverse suite of benchmarks.

It significantly outperforms contemporary vision-language models, including larger networks like OVIS2.5-9B and point-centric specialists such as Rex-Omni-3B. Notably, our model scores 83.9 F1@Point on COCO and exhibits exceptional resilience in heavily packed environments, reaching 87.6 F1@Point on Dense200. Furthermore, it demonstrates superior alignment of complex human intents to spatial regions, achieving 84.7 F1@Point on HumanRef and 91.0 F1@Point on the RefCOCOg test set. These results underscore the effectiveness of our box-aligned training paradigm and the massive scale of LocateAnything-Data in establishing precise geometric alignments, extending seamlessly to point-based generation.

C.2. Comprehensive Performance Across Decoding Modes

In this section, we provide a detailed breakdown of LocateAnything's performance across its three on-demand decoding modes: **Fast**, **Hybrid**, and **Slow**. These modes allow for a dynamic trade-off between geometric

Table 11: Performance evaluation for the object pointing task across a diverse range of benchmarks (COCO, LVIS, Dense200, VisDrone, RefCOCOg, HumanRef). F1-scores are used as the primary metric. The results of the *Hybrid Mode* are reported here.

Method	COCO	LVIS	Dense200	VisDrone	HumanRef	RefCOCOg val	RefCOCOg test
	F1@Point	F1@Point	F1@Point	F1@Point	F1@Point	F1@Point	F1@Point
OVIS2.5-2B	73.4	52.8	36.4	23.8	72.5	83.1	83.1
Qwen2.5-VL-3B	65.9	48.3	4.3	13.9	64.1	77.4	77.8
Qwen2.5-VL-7B	61.1	56.5	2.0	14.2	65.1	78.9	79.4
OVIS2.5-9B	72.6	61.7	35.0	18.8	62.3	85.0	84.5
Molmo-7B-D	77.3	40.3	33.1	29.2	70.0	83.7	83.6
SEED1.5-VL	78.2	70.7	72.1	56.7	83.1	83.6	84.2
Rex-Omni-SFT-3B	76.0	66.7	72.9	49.5	82.1	83.3	83.9
Rex-Omni-3B	80.5	70.8	82.5	58.9	83.8	84.7	85.1
LocateAnything-3B	**83.9**	**76.6**	**87.6**	**60.4**	**84.7**	**91.3**	**91.0**

Table 12: Comprehensive performance of our Fast, Hybrid, and Slow configurations across multiple visual tasks. Throughput (measured in Boxes Per Second, BPS) is reported in the header for each mode. For general detection (COCO, LVIS), we report Average Precision (AP), Average Recall (AR), and F1@mIoU. For other tasks, we report the primary comprehensive metric (e.g., F1@mIoU, F1@mIoU, Avg Acc).

Task Group	**Dataset**	**Metric**	**Fast Model** (15.3 BPS)	**Hybrid Model** (12.7 BPS)	**Slow Model** (4.3 BPS)
Detection	COCO	P@mIoU	58.9	**60.8**	**60.8**
		R@mIoU	46.8	49.7	**50.3**
		F1@mIoU	52.2	54.7	**55.1**
	LVIS	P@mIoU	64.3	**68.4**	68.0
		R@mIoU	37.1	40.3	**42.8**
		F1@mIoU	47.0	50.7	**52.6**
	Dense200	F1@mIoU	46.8	61.3	**61.5**
	VisDrone	F1@mIoU	34.4	39.8	**40.2**
OCR	HierText	F1@mIoU	28.8	29.1	**43.2**
	ICDAR2015	F1@mIoU	26.6	26.4	**27.3**
	TotalText	F1@mIoU	44.4	44.6	**47.5**
	SROIE	F1@mIoU	38.8	39.3	**64.4**
Layout	DocLayNet	F1@mIoU	67.2	77.7	**80.4**
	M6Doc	F1@mIoU	64.1	**70.5**	69.7
GUI	ScreenSpot-Pro	Acc	59.7	60.3	**60.5**
Referring	HumanRef	F1@mIoU	66.8	78.5	**79.1**
	RefCOCOg val	F1@mIoU	70.8	**73.4**	72.4
	RefCOCOg test	F1@mIoU	72.5	**74.8**	73.8
Pointing	COCO	F1@Point	83.1	83.9	**84.8**
	LVIS	F1@Point	74.4	76.6	**76.9**
	Dense200	F1@Point	**89.4**	87.6	88.3
	VisDrone	F1@Point	58.1	60.4	**61.3**

precision and inference latency, as summarized in Tab. 12.

C.3. Backbone Generalization

To examine whether Parallel Box Decoding (PBD) depends on a specific vision-language backbone, we instantiate the same decoding design on Qwen3-VL-4B (Bai et al., 2025a). Following the controlled setting used for the ablation study in the main paper, this variant is trained exclusively on COCO, isolating the effect of the decoding formulation from large-scale data scaling. To examine whether Parallel Box Decoding (PBD) depends on a specific vision-language backbone, we instantiate the same decoding design on Qwen3-VL-4B (Bai et al., 2025a). Following the controlled setting used for the ablation study in the main paper, this variant is trained exclusively on COCO, isolating the effect of the decoding formulation from large-scale data scaling.

As shown in Tab. 13, PBD consistently improves the speed–accuracy trade-off on Qwen3-VL-4B. The Hybrid configuration improves COCO F1 from 50.8 to 52.0 while increasing throughput from 2.8 to 9.4 BPS. These results indicate that the benefits of box-aligned parallel decoding are not tied to a particular backbone architecture.

Table 13: **Backbone generalization.**

Backbone	Decoding	F1	BPS
Qwen3-VL-4B	NTP (baseline)	50.8	2.8
Qwen3-VL-4B	+ PBD (Slow)	**52.2**	2.8
Qwen3-VL-4B	+ PBD (Fast)	49.6	**11.4**
Qwen3-VL-4B	+ PBD (Hybrid)	<u>52.0</u>	<u>9.4</u>

C.4. Mode Analysis and Throughput

Our on-demand decoding modes allow for a dynamic trade-off between geometric precision and inference latency. (1) **Slow Mode (Next-Token Prediction):** Utilizing standard autoregressive generation, this mode consistently establishes the upper bound for localization accuracy (*e.g.*, peak F1@mIoU of 55.1 on COCO and 79.8 on DocLayNet). By processing tokens sequentially, it maintains superior spatial awareness and robust handling of dense object clusters. (2) **Fast Mode (Multi-Token Prediction):** This mode maximizes inference throughput to **15.3 BPS** by predicting full geometric elements in parallel. While it incurs slight accuracy drops in complex or highly dense scenarios, its high-velocity output makes it ideal for real-time applications. (3) **Hybrid Mode (Adaptive Decoding):** Serving as the optimal choice for production pipelines, this mode defaults to parallel decoding and selectively falls back to autoregressive generation only when spatial ambiguity or format irregularities are detected. Operating at a highly competitive **12.7 BPS**, it preserves the speed gains of parallelization while maintaining precise outputs.

C.5. Experimental Setup

To ensure transparency and reproducibility, all performance metrics are reported under specific configurations. For **Throughput Benchmarking**, all values, measured in Boxes Per Second (BPS), were evaluated specifically on the COCO dataset to provide a consistent baseline for speed comparison. Regarding **Input Resolution**, images for the **COCO** and **LVIS** benchmarks were resized with the short side set to 840 pixels. For all other benchmarks, the model was evaluated using the original resolution of the source data.

C.6. Qualitative Comparisons

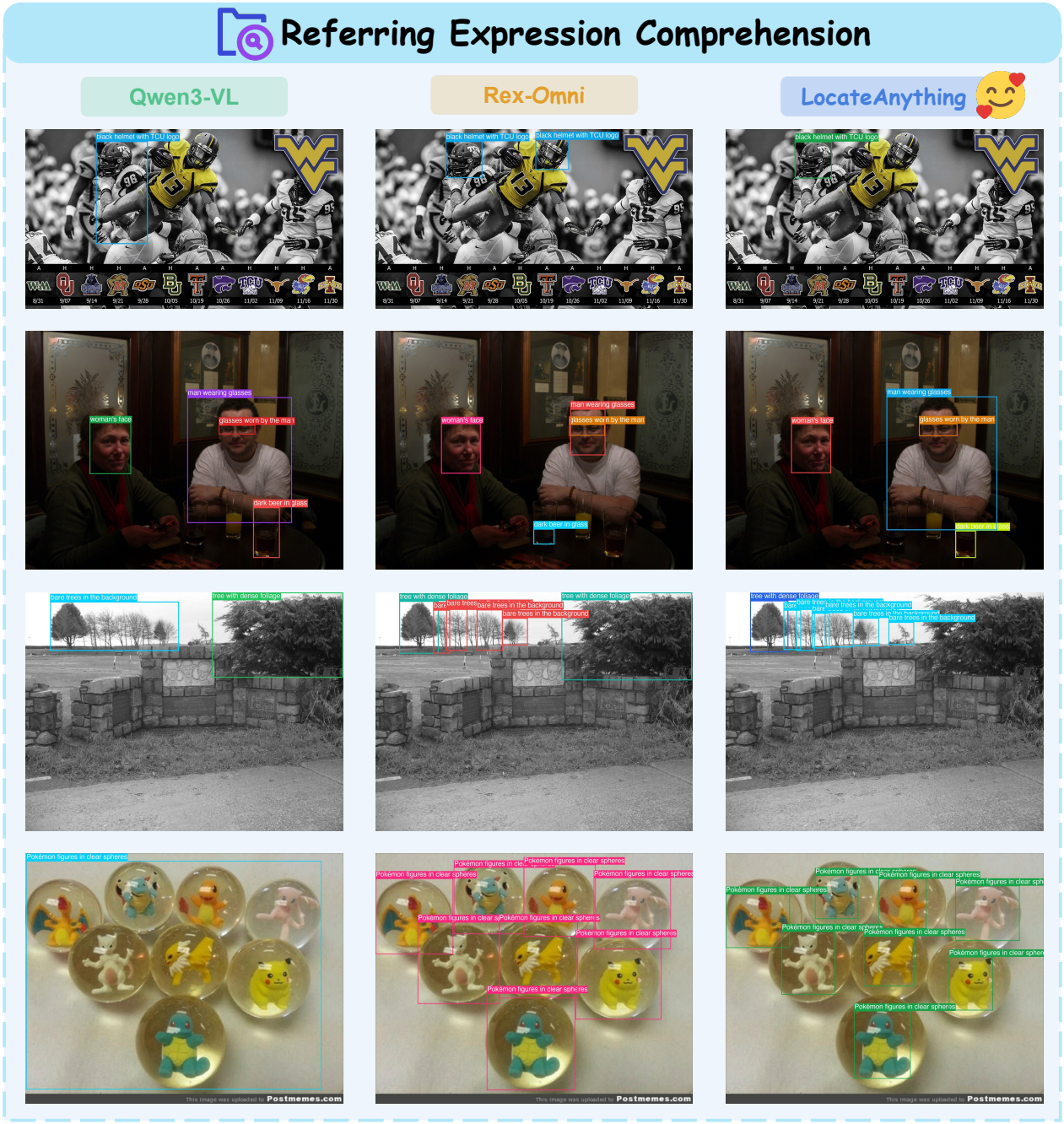


Figure 12: **Qualitative comparison on Referring Expression Comprehension (REC).** Compared to Qwen3-VL and Rex-Omni, LocateAnything demonstrates superior compositional grounding capabilities. It accurately aligns nuanced, free-form human intents (e.g., spatial or attribute-based queries) with correct visual regions.

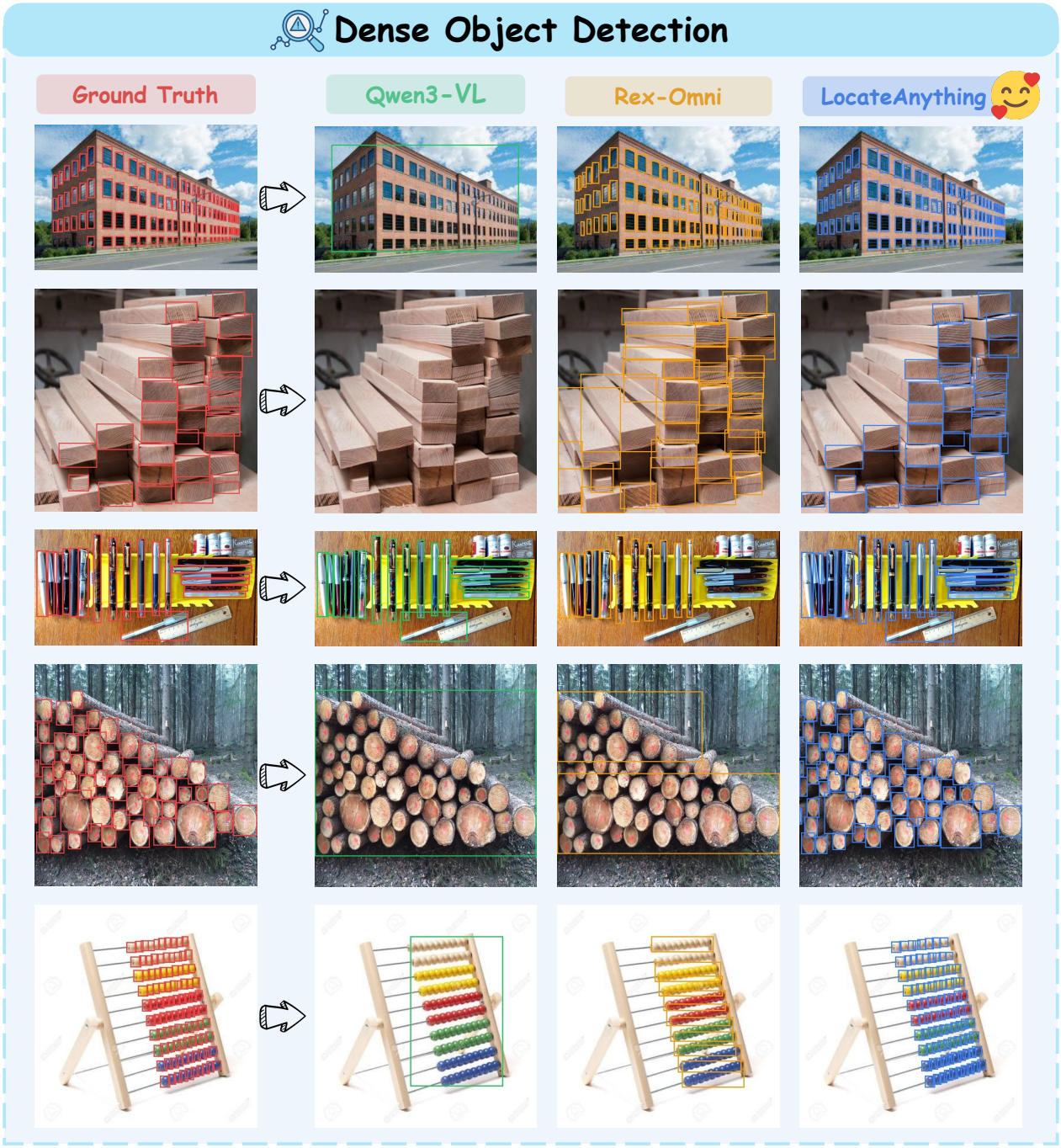


Figure 13: **Qualitative comparison on Dense Object Detection (DOD).** This figure illustrates performance in highly dense and heavily overlapping environments, such as stacked logs and abacus beads. While traditional token-by-token generation models (Qwen3-VL) and point-based models (Rex-Omni) suffer from severe omissions or spatial ambiguity (blurring boundaries between adjacent objects), LocateAnything maintains compact, well-separated, and highly accurate bounding boxes. This confirms the effectiveness of our block-level intra-attention and dense-aware Stage-2 training.

Figure 14: **Qualitative comparison on Optical Character Recognition (OCR).** For scene text (e.g., magazine covers) and structured documents (e.g., tables), LocateAnything yields tightly bounded boxes around text elements. The baseline models frequently exhibit format irregularities or merge distinct text blocks. Our parallel decoding, combined with the Hybrid Mode fallback for complex spatial layouts, ensures high-precision localization without sacrificing structural coherence.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8076–8084, 2019. 15

aidapearson. Aida calculus math handwriting recognition dataset. https://www.kaggle.com/datasets/aidapearson/ocr-data, 2023. 15

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, et al. Block diffusion: Interpolating between autoregressive and diffusion language models. In *ICLR*, 2025. 2, 4, 5, 11

Ashraq. Tmdb-celeb-10k dataset. https://huggingface.co/datasets/ashraq/tmdb-celeb-10k, 2024. 15

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. 15

Azu. Handwritten-mathematical-expression-convert-latex. https://huggingface.co/datasets/Azu/Handwritten-Mathematical-Expression-Convert-LaTeX, 2023. 15

Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025. 2

BAAI. Infinity-instruct dataset. https://huggingface.co/datasets/BAAI/Infinity-Instruct, 2024. 15

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. 3, 8, 9, 10, 17, 18, 19, 22

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. 2, 3

Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*, 2025c. 3

Jonas Belouadi, Anne Lauscher, and Steffen Eger. Automatikz: Text-guided synthesis of scientific vector graphics with tikz. *arXiv preprint arXiv:2310.00367*, 2023. 15

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. 15

Mu Cai, Haotian Liu, Siva Feng, Yong Jae Lee, et al. ViP-LLaVA: Making large multimodal models understand arbitrary visual prompts. In *CVPR*, 2024a. 3

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, et al. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*, 2024b. 2, 3

Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *COLING*, pages 1511–1520, 2022. 15

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, et al. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 2, 3, 8

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719. 18

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. 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. 15

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. *arXiv preprint arXiv:2402.11684*, 2024. 15

Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, et al. EAGLE 2.5: Boosting long-context post-training for frontier vision-language models. *arXiv preprint arXiv:2504.15271*, 2025. 2

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*, 2022a. 15

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Pall Zhu, et al. Shikra: Unleashing multimodal LLM's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023a. 3

Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *ICLR*, 2022b. 2, 4

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, 2023b. 15

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023c. 3

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, et al. SpatialRGPT: Grounded spatial reasoning in vision-language models. *NeurIPS*, 37:135062–135093, 2024. 2

Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, et al. M6Doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *CVPR*, pages 15138–15147, 2023. 9

Chee Kheng Ch'Ng and Chee Seng Chan. Total-Text: A comprehensive dataset for scene text detection and recognition. In *ICDAR*, volume 1, pages 935–942. IEEE, 2017. 9

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. 15

Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In *ACL*, pages 845–855, 2018. 15

LLM-Red-Team Contributors. emo-visual-data: Emotion and visual data analysis project. https://github.com/LLM-Red-Team/emo-visual-data, 2024. 15

Cosmos Team. Nvidia cosmos-reason2. https://huggingface.co/nvidia/Cosmos-Reason2-8B, 2025. 8, 9

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, et al. PaddleOCR 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025. 8, 10

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104, 2025. 15, 18

Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*, pages 845–854, 2017. 15

Patrick Dendorfer, Aljoša Ošep, Anton Milan, Konrad Schindler, Daniel Cremers, et al. MOTChallenge: A benchmark for single-camera multiple target tracking. *arXiv preprint arXiv:2010.07548*, 2020. 8

Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, et al. NVIDIA Nemotron Nano v2 VL. *arXiv preprint arXiv:2511.03929*, 2025. 2

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. 15

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. 15

Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In *ICCVW*, pages 0–0, 2019. 8

EmileEsmaili. sheet music clean ataset. https://huggingface.co/datasets/EmileEsmaili/sheet_music_clean, 2024. 15

FastJobs. Visual emotional analysis dataset. https://huggingface.co/datasets/FastJobs/Visual_Emotional_Analysis, 2024. 15

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. 15

Aarash Feizi, Shravan Nayak, Xiangru Jian, Kevin Qinghong Lin, Kaixin Li, Rabiul Awal, Xing Han Lù, Johan Obando-Ceron, Juan A. Rodriguez, Nicolas Chapados, David Vazquez, Adriana Romero-Soriano, Reihaneh Rabbany, Perouz Taslakian, Christopher Pal, Spandana Gella, and Sai Rajeswar. Grounding computer use agents on human demonstrations, 2025a. URL https://arxiv.org/abs/2511.07332. 17

Aarash Feizi, Shravan Nayak, Xiangru Jian, Kevin Qinghong Lin, Kaixin Li, et al. Grounding computer use agents on human demonstrations. *arXiv preprint arXiv:2511.07332*, 2025b. 2

flytech. Python codes 25k dataset. https://huggingface.co/datasets/flytech/python-codes-25k, 2024. 15

Shenghao Fu, Yukun Su, Fengyun Rao, Jing Xiaohua Lyu, Xie, et al. WeDetect: Fast open-vocabulary object detection as retrieval. *arXiv preprint arXiv:2512.12309*, 2025a. 3

Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, et al. LLMDet: Learning strong open-vocabulary object detectors under the supervision of large language models. *arXiv preprint arXiv:2501.18954*, 2025b. 2, 3

Yonggan Fu, Lexington Whalen, Zhifan Ye, Xin Dong, Shizhe Diao, et al. Efficient-dLM: From autoregressive to diffusion language models, and beyond in speed. *arXiv preprint arXiv:2512.14067*, 2025c. 5

Changlong Gao, Zhangxuan Gu, Yulin Liu, Xinyu Qiu, Shuheng Shen, et al. UI-Venus-1.5 technical report. *arXiv preprint arXiv:2602.09082*, 2026. 8, 9

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Roziere, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. In *ICML*, 2024. 3

Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. Precise detection in densely packed scenes. In *CVPR*, pages 5227–5236, 2019. 8

Shansan Gong, Jiacheng Ye, Zhihui Xie, Zheng Lin, Jiahui Gao, et al. DiffuCoder: Diffusion-based code generation with large language models. *arXiv preprint arXiv:2501.01142*, 2025. 4

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. 15

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, et al. SEED1.5-VL technical report. *arXiv preprint arXiv:2505.07062*, 2025. 8, 9, 10

Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019. 8

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. 15

hazal karakus. mscoco-controlnet-canny-less-colors dataset. https://huggingface.co/datasets/hazal-karakus/mscoco-controlnet-canny-less-colors, 2024. 15

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. 15

Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. Grec: Generalized referring expression comprehension, 2023. URL https://arxiv.org/abs/2308.16182. 17

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. 15

Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2.5: Improved baselines for agglomerative vision foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22487–22497, 2025. 2

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. 15

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. 15

Ailin Huang, Chengyuan Yao, Chunrui Han, Fanqi Wan, Hangyu Guo, et al. Step3-VL-10B technical report. *arXiv preprint arXiv:2601.09668*, 2026. 2

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. 15

HugGAN. Wikiart dataset. https://huggingface.co/datasets/huggan/wikiart, 2024. 15

iamtarun. Python code instructions 18k alpaca dataset. https://huggingface.co/datasets/iamtarun/python_code_instructions_18k_alpaca, 2024. 15

ift. Handwriting forms dataset. https://huggingface.co/datasets/ift/handwriting_forms, 2024. 15

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)*, volume 2, pages 1–6. IEEE, 2019. 15

Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*, 2018. 15

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, et al. Detect anything via next point prediction. *arXiv preprint arXiv:2510.12798*, 2025a. 2, 3, 4, 8, 9, 10, 17, 18

Qing Jiang, Lin Wu, Zhaoyang Zeng, Tianhe Ren, Yuda Xiong, et al. Referring to any person. In *CVPR*, pages 21667–21678, 2025b. 9

Qing Jiang et al. A training-free guess what vision language model from snippets to open-vocabulary object detection. *arXiv preprint arXiv:2601.11910*, 2026. 3

Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, et al. Far3D: Expanding the horizon for surround-view 3D object detection. In *AAAI*, pages 2561–2569, 2024. 3

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, pages 5648–5656, 2018. 15

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. 15

Kamizuru00. Diagram image to text dataset. https://huggingface.co/datasets/Kamizuru00/diagram_image_to_text, 2024. 15

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. 15

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086. URL https://aclanthology.org/D14-1086/. 9

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. 15

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. 15

keremberke. Indoor scene classification dataset. https://huggingface.co/datasets/keremberke/indoor-scene-classification, 2024. 15

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. 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. 15

Kimi Team. Kimi-VL technical report. *arXiv preprint arXiv:2504.07491*, 2025. 4

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. URL https://arxiv.org/abs/2304.02643. 18

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. 15

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 17, 18

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, et al. LISA: Reasoning segmentation via large language model. In *CVPR*, pages 9579–9589, 2024a. 2

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*, 2024b. 15

LAION. Gpt-4v dataset. https://huggingface.co/datasets/laion/gpt4v-dataset, 2023. 15

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. 15

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*, 2024a. 15

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024b. 15

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*, 2024c. 15

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. 15

Jincheng Li, Chunyu Xie, Ji Ao, Dawei Leng, and Yuhui Yin. LMM-Det: Make large multimodal models excel in object detection. In *ICCV*, 2025a. 2

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, et al. ScreenSpot-Pro: GUI grounding for professional high-resolution computer use. In *ACM MM*, pages 8778–8786, 2025b. 9, 10

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*, 2024a. 15

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-LM improves controllable text generation. In *NeurIPS*, 2022. 2

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, et al. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023a. 3

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, 2024b. 15

Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, et al. EAGLE 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025c. 2

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, 2023b. 15

W Lian, B Goodson, E Pentland, et al. Openorca: An open dataset of gpt augmented flan reasoning traces, 2023. 15

Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Zechen Bai, et al. ShowUI: One vision-language-action model for generalist GUI agent. In *NeurIPS Workshop*, 2024a. 2

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, et al. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 8

Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2403.20271*, 2024b. 3

Weifeng Lin, Xinyu Wei, Ruichuan An, Tianhe Ren, Tingwei Chen, Renrui Zhang, Ziyu Guo, Wentao Zhang, Lei Zhang, and Hongsheng Li. Perceive anything: Recognize, explain, caption, and segment anything in images and videos. *arXiv preprint arXiv:2506.05302*, 2025. 3

Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022. 15

Aiwei Liu, Minghua He, Shaoxun Zeng, Sijun Zhang, Linhao Zhang, et al. WeDLM: Reconciling diffusion language models with standard causal attention for fast inference. *arXiv preprint arXiv:2512.22737*, 2025a. 5

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, et al. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a. 2

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*, 2023a. 15

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*, 2023b. 15

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2023c. 15

Jingyu Liu, Xin Dong, Zhifan Ye, Rishabh Mehta, Yonggan Fu, et al. TiDAR: Think in diffusion, talk in autoregression. *arXiv preprint arXiv:2511.08923*, 2025b. 2

Junpeng Liu, Tianyue Ou, Yifan Song, Yuxiao Qu, Wai Lam, Chenyan Xiong, Wenhu Chen, Graham Neubig, and Xiang Yue. Harnessing webpage uis for text-rich visual understanding, 2024b. URL https://arxiv.org/abs/2410.13824. 15

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023d. 2, 8, 9, 10

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*, 2024c. 15

Yangzhou Liu, Yue Cao, Hao Li, Gen Luo, Zhe Chen, et al. Sequential diffusion language models. *arXiv preprint arXiv:2509.24007*, 2025c. 2, 4, 5, 11

Yuhang Liu, Pengxiang Li, Congkai Xie, Xavier Hu, Xiaotian Han, et al. InfiGUI-R1: Advancing multimodal GUI agents from reactive actors to deliberative reasoners. *arXiv preprint arXiv:2504.14239*, 2025d. 8, 9

Zhaoyang Liu, JingJing Xie, Zichen Ding, Zehao Li, Bowen Yang, et al. ScaleCUA: Scaling open-source computer use agents with cross-platform data. *arXiv preprint arXiv:2509.15221*, 2025e. 2, 8, 9

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*, 2024d. 15

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. 15

LooksJuicy. Ruozhiba dataset. https://huggingface.co/datasets/LooksJuicy/ruozhiba, 2024. 15

Guanxi Lu et al. AdaBlock-dLLM: Semantic-aware diffusion LLM inference via adaptive block size. *arXiv preprint arXiv:2509.26432*, 2025a. 4

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Intergps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021a. 15

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Intergps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021b. 15

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*, 2021c. 15

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, 2022a. 15

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*, 2022b. 15

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, et al. Ovis2.5 technical report. *arXiv preprint arXiv:2508.11737*, 2025b. 8, 9, 10

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. 15

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*, 2024a. 15

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. *arXiv preprint arXiv:2404.13013*, 2024b. 3

Divyat Mahajan, Sachin Goyal, Badr Youbi Idrissi, Mohammad Pezeshki, Ioannis Mitliagkas, et al. Beyond multi-token prediction: Pretraining LLMs with future summaries. *arXiv preprint arXiv:2510.14751*, 2025. 3

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. 15

Yunze Man, Shihao Wang, Guowen Zhang, Johan Bjorck, Zhiqi Li, et al. Locateanything3d: Vision-language 3d detection with chain-of-sight. *arXiv preprint arXiv:2511.20648*, 2025. 2

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. 15

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. 15

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. 15

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. 15

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022. 15

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *WACV*, pages 1527–1536, 2020. 15

Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*. BMVA, 2012. 15

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. 15

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. 15

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. 15

mPLUG. Docreason25k dataset. https://huggingface.co/datasets/mPLUG/DocReason25K, 2024. 15

mychen76. Invoices and receipts ocr v1 dataset. https://huggingface.co/datasets/mychen76/invoices-and-receipts_ocr_v1, 2024. 15

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. 15

Shravan Nayak, Xiangru Jian, Kevin Qinghong Lin, Juan A Rodriguez, Montek Kalsi, et al. UI-Vision: A desktop-centric GUI benchmark for visual perception and interaction. *arXiv preprint arXiv:2503.15661*, 2025. 2

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, et al. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025. 2, 3, 5

OleehyO. Latex formulas dataset. https://huggingface.co/datasets/OleehyO/latex-formulas, 2024. 15

OpenGVLab. Sharegpt-4o dataset. https://huggingface.co/datasets/OpenGVLab/ShareGPT-4o, 2024. 15

parasam. Captcha dataset. https://www.kaggle.com/datasets/parsasam/captcha-dataset, 2024. 15

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, et al. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. DocLayNet: A large human-annotated dataset for document-layout segmentation. In *KDD*, pages 3743–3751, 2022. 9

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. 15

PixArt-alpha. Sam-llava-captions10m dataset. https://huggingface.co/datasets/PixArt-alpha/SAM-LLaVA-Captions10M, 2024. 15

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, 2016. URL https://arxiv.org/abs/1505.04870. 17

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. 15

Yu Qi, Yumeng Zhang, Chenting Gong, Xiao Tan, Weiming Zhang, et al. CoT4Det: A chain-of-thought framework for perception-oriented vision-language tasks. *arXiv preprint arXiv:2512.06663*, 2025. 2, 3

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/. 4

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. 15

Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 15

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12490–12500, 2024. 2

Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28, 2015. 15

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6):1137–1149, 2016. 2, 3, 8

Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, et al. Grounding DINO 1.5: Advance the "edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024a. 3

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, et al. PixelLM: Pixel reasoning with large multimodal model. In *CVPR*, pages 26374–26383, 2024b. 2

Mohammad Samragh, Arnav Kundu, David Harrison, Kumari Nishu, Devang Naik, et al. Your LLM knows the future: Uncovering its multi-token prediction potential. *arXiv preprint arXiv:2507.11851*, 2025. 3

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. 15

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. 15

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *AAAI*, pages 8876–8884, 2019. 15

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 17, 18

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 15

shreyanshu09. Block diagram dataset. https://huggingface.co/datasets/shreyanshu09/Block_Diagram, 2024. 15

Chenglei Si, Yanzhe Zhang, Ryan Li, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2code: Benchmarking multimodal code generation for automated front-end engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3956–3974, 2025. 15

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. 15

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. 15

skvarre. Movie posters-100k dataset. https://huggingface.co/datasets/skvarre/movie_posters-100k, 2024. 15

Yongyi Su, Haojie Zhang, Shijie Li, Nanqing Liu, Jingyi Liao, et al. Patch-as-decodable-token: Towards unified multi-modal vision tasks in MLLMs. In *ICLR*, 2026. 3

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. 15

Sujet AI, Allaa Boutaleb, and Hamed Rahimi. Sujet-Finance-QA-Vision-100k: A large-scale dataset for financial document VQA. https://huggingface.co/datasets/sujet-ai/Sujet-Finance-QA-Vision-100k, 2024. 15

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. 15

TAL. Tal open dataset. https://ai.100tal.com/dataset, 2023. 15

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. 15

Benny J Tang, Angie Boggust, and Arvind Satyanarayan. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*, 2023. 15

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. 15

Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants. https://huggingface.co/datasets/teknium/OpenHermes-2.5, 2023. 15

TIGER-Lab. Visualwebinstruct dataset. https://huggingface.co/datasets/TIGER-Lab/VisualWebInstruct, 2024. 15

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. 15

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*, 2024. 15

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. 15

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. 15

VQAonDB. Vqaondb dataset. https://ilocr.iiit.ac.in/vqabd/. 15

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*, 2023a. 15

Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *ICCV*, pages 3621–3631, 2023b. 2

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, et al. InternVL 3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a. 2

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. 15

Xu Wang et al. Diffusion LLMs can do faster-than-AR inference via discrete diffusion forcing. *arXiv preprint arXiv:2508.09192*, 2025b. 4, 5

wendlerc. Renderedtext dataset. https://huggingface.co/datasets/wendlerc/RenderedText, 2024. 15

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. 15

Chengyue Wu, Hao Zhang, Shuchen Xue, Shizhe Diao, Yonggan Fu, et al. Fast-dLLM v2: Efficient block-diffusion LLM. *arXiv preprint arXiv:2509.26328*, 2025a. 4, 5

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, et al. Fast-dLLM: Training-free acceleration of diffusion LLM by enabling KV cache and parallel decoding. *arXiv preprint arXiv:2505.22618*, 2025b. 5

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*, 2024a. 15

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, et al. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024b. 8, 9, 10

Xiaomi Team. MiMo-VL technical report. *arXiv preprint arXiv:2506.03569*, 2025. 8, 9, 10

Tianbao Xie, Jiaqi Deng, Xiaochuan Li, Junlin Yang, Haoyuan Wu, et al. Scaling computer-use grounding via user interface decomposition and synthesis. *arXiv preprint arXiv:2505.13227*, 2025. 8, 9

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. 15

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. 15

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. 15

Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. Pink: Unveiling the power of referential comprehension for multi-modal LLMs. In *CVPR*, 2024. 3

Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, et al. Kwai Keye-VL 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025a. 2

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. 15

Yan Yang, Dongxu Li, Yutong Dai, Yuhao Yang, Ziyang Luo, et al. GTA1: GUI test-time scaling agent. *arXiv preprint arXiv:2507.05791*, 2025b. 8, 9

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. 15

Jiabo Ye, Xi Zhang, Haiyang Xu, Haowei Liu, Junyang Wang, et al. Mobile-Agent-v3: Fundamental agents for GUI automation. *arXiv preprint arXiv:2508.15144*, 2025a. 8, 9

Jiacheng Ye, Zhihui Xie, Zheng Lin, Jiahui Gao, Zirui Wu, et al. Dream 7B: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025b. 2, 3

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, et al. Ferret: Refer and ground anything anywhere at any granularity. In *ICLR*, 2024. 2

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, et al. Perception-R1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025. 2

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions, 2016. URL https://arxiv.org/abs/1608.00272. 17

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*, 2023a. 15

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*, 2024a. 15

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, 2023b. 15

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*, 2024b. 15

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. 15

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. 15

Yuqian Yuan, Wenqiao Zhang, Xin Li, Shihao Wang, Kehan Li, et al. PixelRefer: A unified framework for spatio-temporal object referring with arbitrary granularity. *arXiv preprint arXiv:2510.23603*, 2025. 2

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. 15

Lunbin Zeng, Jingfeng Yao, Bencheng Liao, Hongyuan Tao, Wenyu Liu, et al. DiffusionVL: Translating any autoregressive models into diffusion vision language models. *arXiv preprint arXiv:2512.15713*, 2025. 2, 4

Tao Zewei and Huang Yunpeng. Magiattention: A distributed attention towards linear scalability for ultra-long context, heterogeneous mask training. https://github.com/SandAI-org/MagiAttention/, 2025. 14, 16

Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, et al. Griffon: Spelling out all object locations at any granularity with large language models. In *ECCV*, 2024. 2, 3

Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, et al. Vision-R1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *arXiv preprint arXiv:2503.18013*, 2025. 3

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, 2024a. 15

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, 2019a. 15

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2, 8

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, et al. LLaVA-Grounding: Grounded visual chat with large multimodal models. In *ECCV*, pages 19–35. Springer, 2024b. 2

Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024c. 2, 3

Peng Zhang, Can Li, Liang Qiao, Zhanzhan Cheng, Shiliang Pu, Yi Niu, and Fei Wu. Vsr: a unified framework for document layout analysis combining vision, semantics and relations. In *International conference on document analysis and recognition*, pages 115–130. Springer, 2021. 15

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*, 2024d. 15

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, 2019b. 15

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*, 2024e. 15

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023a. 15

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*, 2023b. 15

Xiangyu Zhao, Shengyuan Ding, Zicheng Zhang, Haian Huang, Maosongcao Maosongcao, Jiaqi Wang, Weiyun Wang, Xinyu Fang, Wenhai Wang, Guangtao Zhai, et al. Omnialign-v: Towards enhanced alignment of mllms with human preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18490–18515, 2025. 15

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. 15

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. 15

Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. DocLayout-YOLO: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *arXiv preprint arXiv:2410.12628*, 2024. 8, 10

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9102–9124, 2024a. 15

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*, 2024b. 15

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. 15

Hanzhang Zhou, Xu Zhang, Panrong Tong, Jianan Zhang, Liangyu Chen, et al. MAI-UI technical report: Real-world centric foundation GUI agents. *arXiv preprint arXiv:2512.22047*, 2025. 8, 9

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. 15

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, et al. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2021. 8