ARGUS: Vision-Centric Reasoning with Grounded Chain-of-Thought

Yunze Man[1,2*], De-An Huang[2], Guilin Liu[2], Shiwei Sheng[2], Shilong Liu[2*]
Liang-Yan Gui[1], Jan Kautz[2], Yu-Xiong Wang[1†], Zhiding Yu[2†]

[1]University of Illinois Urbana-Champaign [2]NVIDIA

Figure 1. Visual question answering, grounding, and chain-of-thought reasoning with Argus. "ctx-token" is short for context token.

Abstract

Recent advances in multimodal large language models (MLLMs) have demonstrated remarkable capabilities in vision-language tasks, yet they often struggle with vision-centric scenarios where precise visual focus is needed for accurate reasoning. In this paper, we introduce Argus to address these limitations with a new visual attention grounding mechanism. Our approach employs object-centric grounding as visual chain-of-thought signals, enabling more effective goal-conditioned visual attention during multimodal reasoning tasks. Evaluations on diverse benchmarks demonstrate that Argus excels in both multimodal reasoning tasks and referring object grounding tasks. Extensive analysis further validates various design choices of Argus, and reveals the effectiveness of explicit language-guided visual region-of-interest engagement in MLLMs, highlighting the importance of advancing multimodal intelligence from a visual-centric perspective.

1. Introduction

Recent breakthroughs in the training of multimodal large language models (MLLMs) [38, 40, 42, 47, 69, 72, 75, 79, 82, 94] have unlocked great advancements in visual-language fusion, allowing these models to extract meaningful content from complex images and perform sophisticated reasoning tasks. However, predominately driven by the success from stronger large language models (LLMs), existing MLLMs still underperform in many vision-centric scenarios [69, 82, 83], where accurate visual perception and understanding determine the success of the subsequent multimodal reasoning tasks (*e.g.,* spatial relationship between objects or properties of specific regions-of-interests (RoIs)). To address these challenges, we *revisit the design space of MLLMs from a vision-centric perspective*, draw insights from cognitive visual intelligence, and propose a new visual attention grounding mechanism for multimodal reasoning tasks, as shown in Figure 1.

Seminal studies in cognitive science [14, 25] have recognized two distinct types of visual attention mechanism, stimulus-driven visual attention and goal-directed visual attention, which are also referred to as involuntary attention and voluntary attention [25, 29, 57, 58], respectively. Stimulus-driven visual attention is an automatic bottom-up capture of attention driven by salient objects or textures in the visual stimulus. On the other hand, goal-directed attention is a top-down conscious selection of attention, driven by goals and intentions. Surprisingly, an interest-

*Work done during an internship at NVIDIA.
†Equal advising.

ing analogy of the two mechanisms of visual attention is presented in the design space of the MLLMs, where (1) the image tokenization with pre-trained visual foundation models [34, 60, 63] represents the stimulus-driven attention, and (2) the language-conditioned image feature engagement happening within the LLM's transformer layers represents the goal-driven attention. The illustration of the visual attention activation strengths in Figure 2 clearly demonstrates different focus areas of two attention modules in MLLMs.

Although several existing methods [69, 82, 104] have studied and highlighted the importance of unconditioned image tokenization to modern MLLMs' reasoning capacity through knowledge distillation and mixture-of-vision-experts (MoVEs), the effect of explicit language-guided visual engagement is underexplored in the research community. This raises two natural questions:

1) *What is the best way to introduce a language-directed visual attention mechanism into the design of MLLMs?*

2) *In addition to perception tasks, can a more explicit visual engagement benefit multimodal reasoning tasks?*

To answer these questions, we explore and propose a grounding-driven visual attention re-engagment module in the multimodal causal prediction process. Unlike most existing MLLMs relying on an implicit self-attention mechanism to model language-directed visual token attendance [47, 82], we pivot to an *explicit top-down visual search* to locate the image RoI most relevant to the text prompt, and then guide the model to focus on the searched regions for subsequent reasoning and answer generation. Recent work has shown that an object-centric representation benefits the vision-language alignment process and subsequent perception tasks [5, 99]. Hence, we utilize text-to-box object-centric grounding as the intermediate reasoning stage, where the predicted bounding boxes serve as simple but effective *visual chain-of-thought (CoT)* signals to help improve the quality of the final reasoning step.

The proposed method, **Argus**, is benchmarked across a diverse set of evaluation datasets, excelling not only in multimodal visual reasoning tasks [24, 53, 54, 71, 82, 83, 93, 94, 102], but also in object-centric visual grounding tasks [30, 100]. We also systematically analyze different designs of the visual attention re-engagement mechanism, and its collaboration with involuntary attention in MLLMs. We hope that our study paves the way towards stronger multimodal intelligence by emphasizing more vision-centric and perception-driven reasoning mechanisms.

2. Related Work

Visual Reasoning with MLLMs. The emergence of multimodal large language models (MLLMs) has revolutionized visual reasoning capabilities, allowing sophisticated question-answering and complex visual understanding tasks. Visual instruction tuning [46] pioneers this advancement by establishing a foundation for tuning language models to handle multimodal tasks effectively. Following this, several improvements and architectural innovations have emerged to enhance zero-shot generalization abilities of MLLMs, including better visual-linguistic alignment [1, 11, 21, 42], high-resolution visual input [38, 47, 88], and dataset curation pipelines [12, 82, 88]. This progress extends to proprietary models like GPT-4V [79], Claude 3 [72], and Gemini [75, 76], which have showcased remarkable general-purpose applicability.

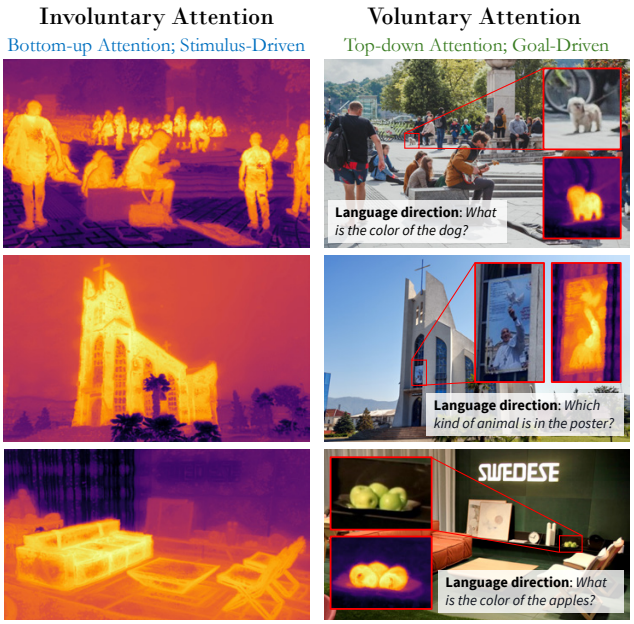


Figure 2. Illustration of two visual attention mechanisms. *Involuntary Attention (Left)*: stimulus-driven; unconditioned feature extraction; salient objects. *Direct Attention (Right)*: Goal-driven; language-guided region-of-interest (RoI) feature extraction.

Several recent studies have shifted focus towards exploring visual reasoning from a vision-centric perspective. Cambrian-1 [82] conducts comprehensive investigations into various visual encoder architectures and introduces a specialized benchmark CV-Bench [82] for assessing vision-centric reasoning capabilities. Eagle [69] further advances this direction by introducing the mixture-of-vision-experts mechanism, demonstrating the potential of specialized visual processing highways in MLLM architectures. However, despite these advancements, current approaches lack conscious control over visual attention mechanisms and do not incorporate explicit goal-driven strategies for visual token extraction, which motivates the introduction of our method.

Visual Perception with MLLMs. Visual perception has consistently been a critical and challenging task within the field of computer vision, encompassing fundamental tasks such as classification, detection, segmentation, and captioning. Numerous specialized models [33, 39, 49, 63, 66], or

"vision experts," have been developed to tackle these tasks. The emergence of MLLMs [46, 79] offers new opportunities to perception tasks. One line of work aims to construct multimodal agents that use MLLMs as controllers to activate specific visual experts [48, 92]. While this framework shows promising performance, it remains complex and unwieldy, limiting its practical applicability. Another line of work involves building unified MLLMs to handle a wide range of vision tasks [89, 95, 99]. Although this approach is feasible in certain scenarios, its overall performance still falls short compared to specialized vision models. Recently, more attention has shifted toward unified models that aspire to be comprehensive, spanning both understanding and generation tasks with the support of vast datasets [73, 74]. Despite these advancements, most efforts have focused less on the synergy between visual perception and reasoning. In contrast, our research explores a grounded CoT approach to vision-centric reasoning, leveraging visual perception as a foundational component.

Chain-of-thought Reasoning. Chain-of-Thought (CoT) reasoning, first introduced by Wei et al. [91], demonstrates that prompting language models to generate intermediate reasoning steps significantly improves their problem-solving capabilities. This concept has sparked numerous works to further enhance reasoning performance, including zero-shot reasoning [35], automatic CoT prompt generation [108], and techniques like self-consistency prompting [90]. Recent works have expanded beyond traditional linear reasoning paths, introducing more sophisticated frameworks such as Graph-of-Thoughts for complex problem decomposition [3], Program of Thoughts for structured numerical reasoning, [9] and Tree of Thoughts for deliberate decision-making [98]. In the multimodal domain, researchers have begun adapting CoT principles to vision-language tasks [22, 41, 52, 67, 109, 111], using off-the-shelf object detectors or multi-turn visual instructions to improve visual reasoning for ambiguous instructions [59, 68, 93, 105], or interleaved segmentation and question answering to develop joint perception and reasoning models [64, 107, 110]. While these works demonstrate the potential of CoT in multimodal contexts, our work, Argus, is the first to systematically study the mechanism of explicit visual attention engagement as visual CoT signals within the MLLM design space, bridging the gap between grounding and vision-centric VQA tasks, while achieving state-of-the-art performance across diverse benchmarks in both domains.

3. Model

In this section, we demonstrate how we design the architecture of Argus from a vision-centric perspective. We start by revisiting the general design space of recent MLLMs [38, 40, 42, 46, 47, 69, 82]. Most existing open-source visual reasoning MLLMs follow a unified autoregressive architectural paradigm, where input images are first converted into visual tokens and concatenated with language tokens before being jointly processed by the LLM for answer token prediction and decoding. This transformation process employs a visual encoder, typically a Vision Transformer (ViT-based [16]) architecture such as CLIP [63], followed by a multilayer perceptron (MLP-based) projector that maps visual features into the text token space. Although this unified architecture has demonstrated strong multimodal capabilities, *it is not optimized for vision-centric scenarios that require precise visual focus for accurate reasoning*. To address this limitation, we propose Argus, a vision-centric reasoning framework with grounded chain-of-thought that enhances MLLMs through explicit language-directed visual region-of-interest (RoI) search and context re-engagement. We present our architectural designs in Section 3.1, and detail the directed visual context re-engagement module in Section 3.2. We validate our design choices and highlight the key empirical findings in Section 4.5. The complete Argus architecture is illustrated in Figure 3.

3.1. Architectural Designs

Visual Encoders. In multimodal vision-centric reasoning, visual encoders play a crucial role by ensuring minimal information loss during image-to-token abstraction and facilitating efficient vision-language alignment. We implement a mixture-of-vision-experts (MoVEs) strategy in our visual encoder suite, building upon recent MLLM studies [69, 82] that demonstrate the complementary benefits of combining different vision foundation models. Our encoding system incorporates three vision experts: CLIP [63], ConvNeXt [50], and EVA-02 [17, 18]. Following extraction, the 2D embeddings are interpolated to uniform spatial dimensions and concatenated along the channel dimension [43, 69]. For multimodal connectors, we employ MLP projectors, consistent with the practice of leading MLLM architectures [38, 40, 47, 69, 82].

The combined visual and language tokens form a multimodal input sequence that is processed by an autoregressive large transformer for next-token prediction. We leverage state-of-the-art pretrained LLMs [13, 78] as our transformer decoder, due to their robust zero-shot reasoning capabilities.

Region-of-Interest Sampling. To enable explicit visual search, we incorporate region-of-interest (RoI) prediction capabilities, allowing our model to output bounding boxes corresponding to regions referenced in the question prompt. This approach is approximately equivalent to the object grounding task, except in the visual reasoning task, our approach extends beyond well-defined objects to handle arbitrary regions relevant to visual reasoning. To maintain a simple design, we adopt the text encoding strategy for

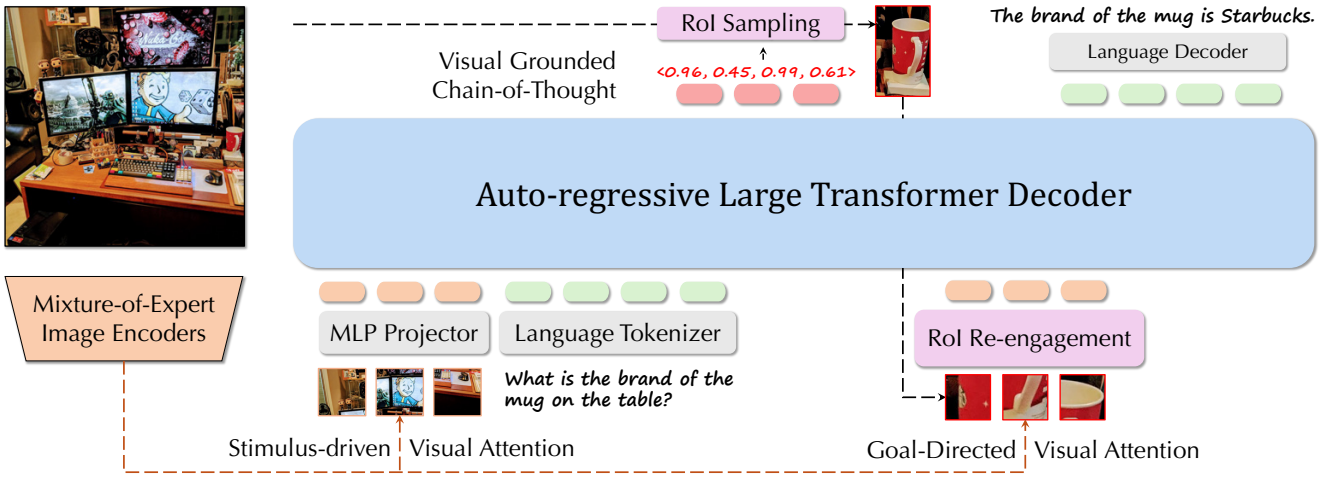


Figure 3. Illustration of Argus architecture. In addition to standard unconditioned visual tokenization, our method incorporates an additional goal-directed visual tokenization procedure. The model has the ability to ground most relevant region-of-interest (RoI) conditioned on the multimodal input instructions. Then, the visual RoI is sampled from the input image, and fed to the RoI re-engagement module to extract another set of visual tokens as CoT context for reasoning.

bounding box representation, where bounding boxes are normalized into $[0, 1]$ range, and represented in text format ($[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$) [7, 99, 104]. This approach eliminates the complexity of training additional box or mask decoding heads [49, 65, 103, 107, 110]. As illustrated in Figure 3, the predicted bounding box guides the cropping process of relevant RoIs from the input image for subsequent visual context re-engagement.

3.2. Directed Visual Context Re-engagement

The model-predicted bounding boxes represent the visual context most relevant to the current reasoning objective. To leverage these regions-of-interest (RoIs) effectively, we seek to direct the model's attention towards these critical areas, thus enhancing focus on context pertinent to the language-defined objectives. Figure 2 illustrates the denoised attention map [96] of our CLIP encoder, showcasing how RoI-specific visual attendance accentuates the essential visual cues aligned with the goal. However, the optimal method for directing this attention remains an underexplored challenge. We identify and categorize four distinct strategies for guiding MLLMs to engage with sampled bounding boxes and incorporate these approaches within the Argus architecture for unified comparison and analysis.

Implicit Self-Attention. Most existing MLLMs do not adopt explicit visual search or attending modules [46, 47, 69, 82]. Instead, they rely on the intrinsic ability of LLMs to attend to visual context through the global self-attention layers. This implicit approach to RoI engagement adopts a minimalistic design, offering simplicity but limited control over specific attention to the bounding boxes.

Implicit Box Guidance. This strategy extends beyond basic self-attention by predicting bounding boxes as special tokens or text coordinates, without explicit visual RoI re-engagement. Although predominantly employed in perception tasks [7, 43, 99, 104], this design can be extended to visual reasoning scenarios, where bounding box predictions serve as chain-of-thought signals, nudging self-attention implicitly towards the RoIs for reasoning purposes. By maintaining the CoT in text format, the shift in attention becomes subtler, merging visual and text-centric cues without explicitly emphasizing visual tokens.

Explicit RoI Re-encoding. In contrast to implicit methods, explicit RoI engagement represents visual CoT signals through actual visual tokens. As demonstrated in Figure 4 (*left*), the re-encoding approach processes sampled image crops through vision encoders for tokenization [68] after processing. The processing is equivalent to an augmentation process, which involves square padding of cropped regions to $\max\{\text{width}, \text{height}\}$ region, context expansion with margins, and dimension-specific resizing for vision experts. These tokens are appended to the input sequence, introducing a supplementary visual context that guides reasoning through explicit, context-specific signals. This approach ensures that RoIs are attended to precisely, albeit with an increase in computational requirements due to the additional encoding process.

Explicit RoI Re-sampling. The re-sampling method offers an alternative explicit engagement strategy that reduces computational overhead. As shown in Figure 4 (*right*), rather than treating RoI boxes as new images, re-sampling utilizes visual embeddings from a memory bank [4, 28, 41]. In visual reasoning tasks, tokens are retrieved from the initial MoVE encoder suite and reused as needed. We calculate the overlap between the RoI bounding boxes and the patch embeddings after visual encoders, and resample the

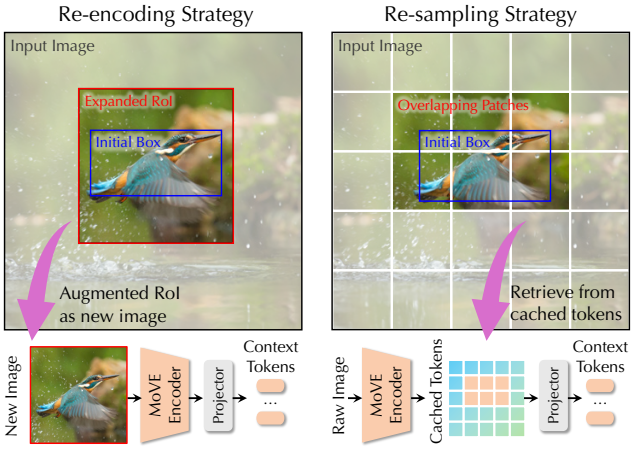


Figure 4. Illustration of two visual CoT mechanisms. Re-encoding expand the RoI and treat it as a new image for tokenization. Re-sampling retrieves knowledge from the pre-extracted token cache.

patch tokens that have intersection with boxes as context tokens for re-engagement. This strategy leverages cached tokens, thus streamlining computation while maintaining a focus on task-relevant regions. In the meanwhile, the redundant tokens also preserve the positional context within the original image, which might be lost during padding and resizing process in the re-encoding method.

4. Experiments

This section presents our comprehensive experimental methodology and results. We begin by detailing our training protocols (Section 4.1), followed by implementation details (Section 4.2). We then describe our evaluation benchmarks and baseline comparisons (Section 4.3), and demonstrate the effectiveness of our model on vision reasoning and reference expression grounding tasks (Section 4.4). Through extensive ablation studies, we validate the design choices of Argus in controlled settings (Section 4.5). More results and details are provided in the supplementary material.

4.1. Training Pipeline

Following the advances of recent MLLMs [42, 46, 47, 69, 82], we divide the training into two stages.

Alignment and Pre-training. In the initial pre-training stage, we adopt the LLaVA-595K dataset [46], which comprises carefully curated image-text pairs. We freeze the LLM while allowing the vision encoders and MLP projector layers to be trainable. Drawing insights from Eagle [69], we implement a vision expert pre-alignment process to minimize representation disparities between experts and enhance the subsequent language alignment.

Supervised Fine-Tuning (SFT). The second stage employs a diverse combination of datasets to ensure robust performance across multiple domains. To ensure ensures strong general-purpose multimodal understanding capabilities, we utilize *Eagle1.8M dataset* [69], a comprehensive collection of conversational data aggregated from various sources [8, 20, 26, 31, 32, 45, 46, 53, 54, 77, 81, 86, 106, 112]. For visual chain-of-thought reasoning, we incorporate the *VCoT dataset* [68], which provides region-of-interest (RoI) bounding box annotations specifically designed for grounding and reasoning tasks. This dataset obtains samples from multiple established benchmarks [23, 24, 37, 44, 54, 55, 62, 70, 71, 84, 85, 112]. We structure each sample as a multi-turn conversation between user and AI agent, where (1) The agent first predicts the region-of-interest using `<roi-box>` annotations in normalized text coordinates (Section 3.1). (2) The user then provides intermediate visual chain-of-thought signals through `<visual-context>` tokens. And (3) the agent generates the final response based on this structured interaction. To enhance our model's ability to ground concepts in unconstrained scenarios, we follow existing work [7, 68, 99] by incorporating a mixture of GRIT [61] (756K) and Shikra [7, 36, 62, 100, 112] (326K) datasets. All spatial grounding information is normalized to the range $[0, 1]$ relative to image dimensions and represented in text format. During this fine-tuning stage, we allow full parameter updates across the MoVE vision encoder, MLP projectors, and the LLM decoder.

4.2. Implementation Details

We use Llama3-8B [78] as our LLM decoder backbone. For vision encoders, we use ViT-L/14 CLIP [63], ConvNeXt-XXL-1024 [50], and EVA-02-L/16 [18] as our MoVE encoding system. The input resolution is set to 1024×1024 for ConvNeXt and EVA-02, and 448×448 for CLIP model. The visual token number is 1024 (32×32). Following Eagle [69], we name our model Argus-X3, to reflect the usage of three vision experts. During the RoI selection, we encode RoI with the format of box coordinates: $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$, and the model is instruction-tuned to directly output them in text, where the numbers are normalized to $[0, 1]$ by image dimensions and accurate to three decimal places. We parse the coordinates by removing brackets and commas, then convert the numbers back to box coordinates for grounding and resampling. For both stages, we train for one epoch with a batch size of 256. The learning rate is set to $1e$-3 for the pre-training stage, and $2e$-5 for the SFT stage. The AdamW optimizer [51] with zero weight decay and a cosine learning rate scheduler is used. Experiments are conducted using NVIDIA A100 GPUs. More details are provided in the supplementary material.

4.3. Baseline Models and Benchmarks

We compare Argus with state-of-the-art MLLMs with roughly the same parameter size, including Mini-Gemini-HD [40], LLaVA-NeXT [47], VisCoT [68], QwenVL [1],

Model	Vision-Centric Tasks						Text Understanding					General Tasks				
Method	Avg	V-Star	CV-Bench^{2D}	CV-Bench^{3D}	MMVP	RealworldQA	Avg	ChartQA	OCRBench	TextVQA	DocVQA	Avg	MMMU^{V}	MMB	SEED^{I}	GQA
ref: GPT-4o (*250128*) [79]	73.7	70.7	79.8	84.3	58.5	75.4	83.1	86.9	75.1	79.7	90.8	-	68.9	87.1	81.7	-
ref: Qwen2.5-VL [2]	72.6	72.8	80.0	82.7	53.1	74.3	87.9	85.2	85.9	84.9	95.7	-	58.6	86.5	78.3	-
ref: InternVL2.5 [80]	71.1	69.1	79.7	81.5	54.9	70.1	84.8	84.8	82.2	79.1	93.0	-	56.0	84.6	79.2	-
7B & 8B Open MLLMs, and Visual-CoT Variants																
Vanilla CoT Prompting	60.3	64.9	66.1	63.5	44.2	62.8	68.2	70.9	56.7	70.8	74.2	62.5	40.3	69.7	75.0	64.8
Visual-CoT-7B [68]	54.4	49.7	61.5	62.4	35.7	62.9	66.6	69.7	51.6	70.0	75.1	60.2	37.2	67.3	74.1	62.0
Mini-Gemini-HD-8B [40]	51.8	52.9	62.2	63.0	18.7	62.1	62.9	59.1	47.7	70.2	74.6	61.9	37.3	72.7	73.2	64.5
LLaVA-NeXT-8B [47]	55.4	50.8	62.2	65.3	38.7	60.1	63.9	69.5	49.0	64.6	72.6	62.9	41.7	72.1	72.7	65.2
InternVL [12]	52.5	52.4	59.2	54.6	36.0	60.4	62.5	70.1	51.1	61.5	67.3	57.6	35.3	64.1	66.9	63.9
QwenVL-7B [1]	52.0	54.5	57.5	55.9	33.3	58.8	60.9	66.3	50.6	61.5	65.1	56.5	35.9	67.0	65.4	57.5
Eagle-X3-8B [69]	59.6	60.7	66.4	63.0	45.1	62.9	67.8	70.4	56.1	70.9	73.9	62.4	39.8	70.9	73.9	64.9
Argus-X3-8B (ours)	**62.2**	**68.1**	**68.5**	64.2	**45.5**	**64.6**	**70.1**	**74.8**	**56.7**	**73.6**	**75.4**	**63.4**	40.4	**72.9**	**75.8**	65.1

Table 1. Argus achieve state-of-the-art performance among public MLLMs of comparable parameter count and training scale.

InternVL [12], and Eagle [69]. We use open-sourced Eagle-X3-8B [69] with the same MoVE encoder structure as our baseline architecture. For reference, we also include performance metrics from proprietary models [79] or models trained with orders of magnitude more (or even undisclosed) data [2, 80]. We use the official evaluation metrics provided by the benchmarks if available. Otherwise, we follow the same evaluation setting as the recent MLLMs [69] for a fair comparison. For grounding tasks, we benchmark against both specialist and generalist models including MAttNet [101], TransVG [15], UNITER [10], VILLA [19], UniTAB [97], MDETR [27], G-DINO [49], OFA-L [87], Shikra [7], Ferret [99], MiniGPT-v2 [6], InternVL2 [80], and Qwen-VL [1]. We report Acc@0.5 as the metrics for the referring expression grounding task. All model performances are obtained from official public reports.

Benchmarks. We report the performance of the methods on various vision-language benchmarks [24, 32, 53, 54, 56, 64, 71, 82, 83, 93, 94, 102]. We follow prior work [68, 69, 82] and combine their evaluation benchmarks to cover a wide spectrum of *vision-centric multimodal* evaluation settings where the keys to correctly answering questions lie in accurate visual understanding.

4.4. Main Performance

Our experimental evaluation focuses on two primary capabilities: *visual reasoning* and *referring expression grounding*. These complementary tasks assess different aspects of our model's multimodal understanding: visual reasoning examines comprehensive multimodal comprehension, while referring expression grounding evaluates precise vision-text alignment through localization tasks.

Performance on Visual Reasoning. We conducted evaluations across three categories of MLLM benchmarks: *General Multimodal Reasoning*, *Text-centric Understanding*, and *Vision-centric Perception*. Performance is shown in Table 1. We observe several findings from the results. (1) First, Argus achieve state-of-the-art performance among public MLLMs with a comparable parameter count and training scale. Notably, our approach even surpasses several proprietary MLLMs, demonstrating its exceptional multimodal reasoning capabilities. (2) Furthermore, we observe substantial improvements in both vision-centric and text understanding tasks. These tasks typically require precise attention to specific visual elements, such as objects or textual components within images, to generate accurate responses. Improvement in these areas highlights the effectiveness of our goal-conditioned visual search mechanism and enhanced visual attention engagement strategies.

Performance on Referring Grounding. To evaluate our model's object grounding capabilities, we utilized the RefCOCO, RefCOCO+, and RefCOCOg benchmarks [100]. The results are shown in Table 2. Our method achieves leading performance among comparable-scale generalist MLLMs, highlighting its effectiveness in combining general-purpose reasoning with precise visual grounding. Our performance is competitive against Grounding-DINO-L [49], a specialist model trained on a larger grounding-specific dataset and optimized for detection tasks. These results demonstrate that Argus not only excels in high-level reasoning tasks, but also achieves exceptional performance in fundamental visual perception and localization tasks.

Qualitative Results. In Figure 5 we demonstrate some qualitative results of Argus on the visually grounded CoT task and the referring grounding task. Our method is capable of achieving complicated visual reasoning tasks with the help of the visually grounded CoT mechanism.

Model	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
Specialist Models								
MAttNet[101]	76.4	80.4	69.3	64.9	70.3	56.0	66.7	67.0
TransVG [15]	81.0	82.7	78.3	64.8	70.7	56.9	68.7	67.7
UNITER [10]	81.4	87.0	74.2	75.9	81.5	66.7	74.0	68.7
VILLA [19]	82.4	87.5	74.8	76.2	81.5	66.8	76.2	76.7
UniTAB [97]	86.3	88.8	80.6	78.7	83.2	69.5	80.0	80.0
MDETR [27]	86.8	89.6	81.4	79.5	84.1	70.6	81.6	80.9
G-DINO [49]	90.6	93.2	88.2	82.8	89.0	75.9	86.1	87.0
Generalist MLLMs								
OFA-L [87]	80.0	83.7	76.4	68.3	76.0	61.8	67.6	67.6
Shikra-7B [7]	87.0	90.6	80.2	81.6	87.4	72.1	82.2	82.2
Ferret-7B [99]	87.5	91.4	82.5	80.8	87.4	73.1	83.9	84.8
MiniGPT-7B [6]	88.7	91.7	85.3	80.0	85.1	74.5	84.4	84.7
InternVL2-8B [80]	87.1	91.1	80.7	79.8	87.9	71.4	82.7	82.7
QwenVL-7B [1]	89.4	92.3	85.3	83.1	88.3	77.2	85.6	85.5
Argus-X3-8B	**89.8**	**92.9**	**85.4**	**84.7**	**90.1**	77.1	**86.7**	85.2

Table 2. Our method achieves leading performance among generalist MLLMs of comparable scale, highlighting its effectiveness in general-purpose reasoning with precise visual grounding.

4.5. Ablation Study and Analysis

The current landscape of the MLLM community presents great challenges for perfectly fair quantitative evaluation and comparison on benchmarks, due to variations in data scale, model sizes, and architectural choices. Hence, in this section, we focus on rigorous and controlled experiments to validate our architectural designs and describe our key findings. We employ an accelerated and unified training schedule across all ablation experiments for fair comparison.

4.5.1 Visual Attention Re-engagement Analysis.

We conducted a systematic evaluation of different visual context re-engagement mechanisms, with results presented in Table 3. (1) First, **the incorporation of CoT reasoning consistently enhances performance across both visual and text-based reasoning tasks**. The introduction of the implicit bounding box CoT guidance yields substantial improvements over implicit attention mechanisms, with explicit CoT reasoning providing even greater performance gains. (2) When comparing re-encoding and re-sampling strategies, we find that **re-sampling generally demonstrates superior performance across most benchmarks**. This advantage can be attributed to the preservation of contextual positional information and the avoidance of distribution shifts that typically occur during region rescaling. However, *this pattern shows an interesting exception in the V-Star benchmark* [93], which emphasizes the perception of small objects in complex scenes. In this specific context, re-encoding proves more effective, as it processes regions with larger patches, thereby preserving more fine-grained details and minimizing information loss.

We make further discussion over the selection of re-encoding and re-sampling strategies. From a high-level perspective, re-sampling involves retrieving tokens from a token cache that was pre-extracted in the initial unconditioned phase. Conversely, re-encoding augments the cropped patch and produces a new embedding, leveraging higher resolution and enhanced contextual focus.

Method	V-Star	CV-Bench-2D	TextVQA	ChartQA
Implicit Att.	58.6	64.5	69.2	67.3
Box Guidance.	63.9	67.0	71.6	70.4
RoI Re-enc.	**68.1**	67.4	71.4	71.8
RoI Re-smp.	67.0	**68.2**	**73.9**	**72.7**

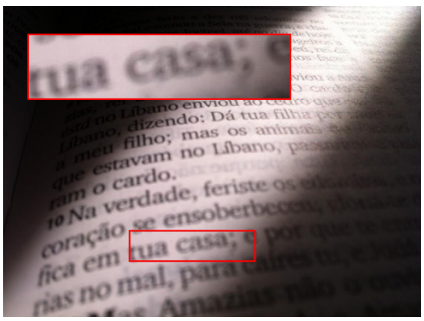
Table 3. Comparison of four visual attention re-engagement mechanisms, as introduced in Section 3.2, respectively. Two explicit visual RoI re-engagement mechanism leads the performance.

Method	V-Star	CV-Bench-2D	TextVQA	ChartQA
Re-encoding				
Δ fewer encoder	–4.2	–3.8	–4.7	–3.9
Δ square context	+1.6	+0.8	+0.1	+2.6
Re-sampling				
Δ fewer encoder	–11.5	–6.1	–4.6	–5.3
Δ square context	+0.5	+0.1	–3.1	1.9
Δ non-share MLPs	0	+1.7	+0.5	+0.3

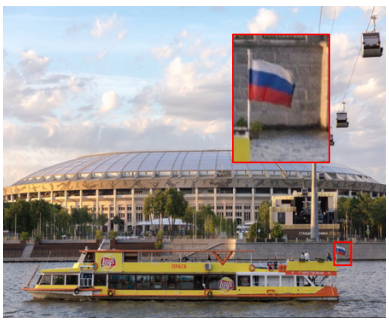
Table 4. Performance change (Δ) after changing design choices between two types of explicit visual re-engagemment strategies. We use red for decrease ↓ and green for increase ↑.

Impact of Visual Encoder Capacity. We begin by examining how re-encoding and re-sampling strategies perform with visual encoders of varying capacities, as illustrated in Table 4 (Δ fewer encoder). We specifically evaluate performance changes when MoVE is replaced with a single CLIP encoder. We have two key observations: (1) **Higher-capacity vision encoders consistently yield improved performance**, which aligns with our expectations that a robust feature extractor enhances subsequent perception and reasoning tasks. (2) **Re-encoding depends less on a high-quality initial feature extraction compared to re-sampling.** Since re-sampling simply retrieves tokens from the cache, it does not have the opportunity to update visual token quality. If detail is lost during the initial extraction, re-sampling struggles to recover it, whereas re-encoding can more effectively refine the visual representation.

Effect of Padding and Square Context. The model's predicted RoI bounding boxes are typically rectangular, and preprocessing is required to convert them to square shapes, as most ViT-based vision encoders recommend. Table 4 (Δ square context) shows how the re-encoding and re-sampling strategies respond to two different preprocessing methods. The default approach pads the image to make it square, while an alternative method expands the RoI to a square based on the larger dimension, effectively capturing more contextual information around the original focus area. Our results indicate that (1) **re-encoding consistently benefits from the larger context introduced by square bounding boxes**, while (2) re-sampling shows only marginal improvement in vision-centric reasoning and a noticeable drop in text-centric (OCR) reasoning. This outcome likely reflects



Question: what is the 3 letter word to the left of casa in the text?
Argus w/o CoT Reasoning: cas ✗
Argus w/ CoT Reasoning:
<roi-box> → <context> → tua ✓



Question: Does the flag have two or three colors?
Argus w/o CoT Reasoning: two colors ✗
Argus w/ CoT Reasoning:
<roi-box> → <context> → three colors ✓

Question:
What is the location of the guy in the black jacket?
What is the location of the guy swinging?
Argus: <box_1> <box_2>

Figure 5. Qualitative evaluation of Argus. We achieve superior performance in challenging multimodal reasoning and perception tasks.

Method	V-Star	CVB-2D	TextVQA	ChartQA
Baseline (Eagle-X3)	55.3	64.9	66.3	63.0
+ *CoT signals*	62.7	65.5	71.1	69.4
++ *Grounding* (Argus)	67.0	68.2	73.9	72.7

Table 5. Impact of CoT and grounding on reasoning performance. We verify that the combination of CoT mechanism and grounding task both benefits the model reasoning capability.

Modules	GMACs	visual tokens	time$^{\text{inference}}$
Re-encoding	8,710.6	1024	827 ms
Re-sampling	**4,355.3**	**26**	**492** ms

Table 6. The re-sampling strategy shows advantages in computational efficiency over the re-encoding strategy.

the nature of text-centric tasks, which often involve locating specific words or sentences within chunks of text (as shown in Figure 5 (*left*)). In these cases, an expanded region may complicate localization tasks by introducing task-irrelevant context, reducing the benefits of bounding box prediction.

Non-shared MLP Layers. For the re-sampling strategy, we also explore the use of non-shared MLP layers as an alternative to the default shared MLP configuration. This involves training a dedicated MLP layer specifically for the RoI re-engagement module. Results shown in Table 4 (Δ non-share MLPs) suggest that separating MLP layers marginally improves performance. We attribute this improvement to the ability of the non-shared MLPs to account for distinct image distributions: one MLP is optimized for raw images with full context, while the other focuses on localized, object-centric regions. This approach effectively combines elements of both re-sampling and re-encoding, resulting in the best overall performance.

4.5.2 Grounding and Reasoning.

We posit that visually grounded CoT directly connects the grounding and reasoning processes. In Table 5, we analyze the impact of CoT signaling and grounding on reasoning performance, with two main observations: (1) By introducing the CoT dataset and re-engagement mechanism into SFT training, our method shows marked improvement in both vision- and text-centric reasoning tasks compared to the Eagle-X3, which is trained on a standard vision-language reasoning dataset, relying on implicit self-attention for attending to visual token. This highlights the importance of CoT for MLLMs. (2) Our full model, Argus, further incorporates grounding datasets into SFT training and exhibits additional performance gains. This enhancement can be attributed to its strengthened object-centric perception, which improves bounding box predictions and, in turn, maximizes the utility of the CoT mechanism.

4.5.3 Computational Efficiency.

We compare the computational overheads introduced by re-encoding and re-sampling strategies. Table 6 shows that re-sampling has a major efficiency advantage in *average visual encoding operations*, measured in giga-multiply-accumulate operations (GMACs), and *number of additional visual tokens*, due to the reuse of patch embeddings, leading to a faster inference time, as LLM prediction is not blocked by visual encoding.

5. Conclusion

This work introduces Argus, a vision-centric reasoning model with grounded chain-of-thought capability. By incorporating a grounding-driven visual attention re-engagement mechanism, Argus demonstrates an effective approach to enhancing multimodal reasoning by emphasizing directed visual focus. Through extensive evaluations, we show that our framework demonstrates superior performance across both multimodal reasoning and referral object grounding tasks. These findings not only advance our understanding of vision-language fusion, but also suggest a promising direction for future MLLM architectures that emphasize vision-centric mechanisms and visual chain-of-thought as a crucial component of multimodal intelligence.

Acknowledgments. We thank Subhashree Radhakrishnan, Fuxiao Liu, and Min Shi for the fruitful discussion on the idea of the project and the implementation of the codebase. Y. Man is supported by the NVIDIA Graduate Fellowship. Y. Man and Y.-X. Wang are supported in part by NSF Grant 2106825 and NIFA Award 2020-67021-32799.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2, 5, 6, 7

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6

[3] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. In *AAAI*, 2024. 3

[4] Nadine Chang, Zhiding Yu, Yu-Xiong Wang, Animashree Anandkumar, Sanja Fidler, and Jose M Alvarez. Image-level or object-level? A tale of two resampling strategies for long-tailed detection. In *ICML*, 2021. 4

[5] Hong-You Chen, Zhengfeng Lai, Haotian Zhang, Xinze Wang, Marcin Eichner, Keen You, Meng Cao, Bowen Zhang, Yinfei Yang, and Zhe Gan. Contrastive localized language-image pre-training. *arXiv preprint arXiv:2410.02746*, 2024. 2

[6] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 6, 7

[7] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal LLM's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 4, 5, 6, 7

[8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4v: Improving large multi-modal models with better captions. In *ECCV*, 2024. 5

[9] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *TMLR*, 2023. 3

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *ECCV*, 2020. 6, 7

[11] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 2, 6

[13] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* chatgpt quality. *https://vicuna. lmsys. org*, 2023. 3

[14] Maurizio Corbetta and Gordon L Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 2002. 1

[15] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-end visual grounding with transformers. In *ICCV*, 2021. 6, 7

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[17] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 3

[18] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 2024. 3, 5

[19] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. 6, 7

[20] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. G-LLaVA: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023. 5

[21] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Jifeng Dai, and

Wenhai Wang. Mini-InternVL: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. *Visual Intelligence*, 2(1):1–17, 2024. 2

[22] Jiaxin Ge, Hongyin Luo, Siyuan Qian, Yulu Gan, Jie Fu, and Shanghang Zhang. Chain of thought prompt tuning in vision language models. *arXiv preprint arXiv:2304.07919*, 2023. 3

[23] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. ICDAR2019 competition on scanned receipt ocr and information extraction. In *ICDAR*, 2019. 5

[24] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 2, 5, 6

[25] William James. *Psychology, briefer course*. Harvard University Press, 1984. 1

[26] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. DVQA: Understanding data visualizations via question answering. In *CVPR*, 2018. 5

[27] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 6, 7

[28] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 4

[29] Stephen Kaplan and Marc G Berman. Directed attention as a common resource for executive functioning and self-regulation. *Perspectives on psychological science*, 2010. 1

[30] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 2

[31] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016. 5

[32] 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. 5, 6

[33] 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. *arXiv:2304.02643*, 2023. 2

[34] 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. In *ICCV*, 2023. 2

[35] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022. 3

[36] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017. 5

[37] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 128(7):1956–1981, 2020. 5

[38] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2, 3

[39] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2

[40] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-Gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 1, 3, 5, 6

[41] Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, and Zhongyu Wei. VoCoT: Unleashing visually grounded multi-step reasoning in large multi-modal models. *arXiv preprint arXiv:2405.16919*, 2024. 3, 4

[42] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. In *CVPR*, 2024. 1, 2, 3, 5

[43] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. SPHINX: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 3, 4

[44] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 5

[45] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 5

[46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 3, 4, 5

[47] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024. 1, 2, 3, 4, 5, 6

[48] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. LLaVA-Plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*, 2023. 3

[49] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Mar-

rying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 2, 4, 6, 7

[50] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 3, 5

[51] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[52] 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. In *NeurIPS*, 2022. 3

[53] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, 2022. 2, 5, 6

[54] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021. 2, 5, 6

[55] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, 2022. 5

[56] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. MM1: Methods, analysis & insights from multimodal LLM pre-training. *arXiv preprint arXiv:2403.09611*, 2024. 6

[57] Robert J Morecraft, Changiz Geula, and M-Marsel Mesulam. Architecture of connectivity within a cingulo-fronto-parietal neurocognitive network for directed attention. *Archives of neurology*, 1993. 1

[58] Vernon B Mountcastle. Brain mechanisms for directed attention. *Journal of the Royal Society of Medicine*, 1978. 1

[59] Minheng Ni, Yutao Fan, Lei Zhang, and Wangmeng Zuo. Visual-O1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning. *arXiv preprint arXiv:2410.03321*, 2024. 3

[60] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 2

[61] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 5

[62] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 5

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 5

[64] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. GLaMM: Pixel grounding large multimodal model. In *CVPR*, 2024. 3, 6

[65] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, Yuda Xiong, Hao Zhang, Feng Li, Peijun Tang, Kent Yu, and Lei Zhang. Grounding DINO 1.5: Advance the edge of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024. 4

[66] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 2

[67] Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*, 2023. 3

[68] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *NeurIPS*, 2024. 3, 4, 5, 6

[69] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the design space for multimodal llms with mixture of encoders. In *ICLR*, 2025. 1, 2, 3, 4, 5, 6

[70] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, 2020. 5

[71] 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*, 2019. 2, 5, 6

[72] Anthropic Team. Introducing the next generation of claude, 2024. 1, 2

[73] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 3

[74] Emu3 Team. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 3

[75] Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2

[76] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2

[77] LAION Team. Laion-gpt4v dataset, 2023. 5

[78] Meta Team. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3, 5

[79] OpenAI Team. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2, 3, 6

[80] OpenGVLab Team. InternVL2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024. 6, 7

[81] Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants., 2023. 5

[82] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 1, 2, 3, 4, 5, 6

[83] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *CVPR*, 2024. 1, 2, 6

[84] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickael Coustaty, Bertrand Anckaert, Ernest Valveny, Matthew Blaschko, Sien Moens, and Tomasz Stanislawek. Document understanding dataset and evaluation (dude). In *ICCV*, 2023. 5

[85] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5

[86] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting GPT-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023. 5

[87] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022. 6, 7

[88] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[89] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. CogVLM: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 3

[90] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023. 3

[91] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. 3

[92] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual ChatGPT: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 3

[93] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *CVPR*, 2024. 2, 3, 6, 7

[94] xAI Team. Grok, 2024. 1, 2, 6

[95] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*, 2024. 3

[96] Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers. In *ECCV*, 2024. 4

[97] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. UniTAB: Unifying text and box outputs for grounded vision-language modeling. In *ECCV*, 2022. 6, 7

[98] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, 2024. 3

[99] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *ICLR*, 2023. 2, 3, 4, 5, 6, 7

[100] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 2, 5, 6

[101] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018. 6, 7

[102] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 2, 6

[103] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, and Jianwei Yang. Llava-grounding:

Grounded visual chat with large multimodal models. In *ECCV*, 2024. 4

[104] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, and Yinfei Yang. Ferret-v2: An improved baseline for referring and grounding with large language models. In *COLM*, 2024. 2, 4

[105] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024. 3

[106] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. LLaVAR: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 5

[107] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. GroundHog: Grounding large language models to holistic segmentation. In *CVPR*, 2024. 3, 4

[108] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *ICLR*, 2023. 3

[109] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 3

[110] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. PSALM: Pixelwise segmentation with large multi-modal model. In *ECCV*, 2024. 3, 4

[111] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. DdCoT: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *NeurIPS*, 2023. 3

[112] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016. 5