# COMPACT CHAIN-OF-THOUGHT DISTILLATION WITH VISUAL GROUNDING GUIDANCE FOR VISION-LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multimodal chain-of-thought (CoT) reasoning has significantly improved the performance of large vision-language models (VLMs) in solving complex tasks. However, distilling this reasoning capability into small VLMs remains challenging. In this paper, we construct a large-scale multimodal CoT dataset comprising 120k samples, each featuring a compact CoT with a standardized structure, explicit length constraints, and region-level visual grounding annotations. Based on our dataset, we obtain two critical observations: (1) VLM layers vary substantially in their receptiveness to visual grounding knowledge; (2) Compact CoTs achieve greater reasoning efficiency and higher accuracy than lengthy native CoTs. Motivated by this, we propose the Visual Grounding-Guided CoT (VGG-CoT) distillation framework, which first identifies receptive layers via per-layer probing and restricts LoRA updates to these layers for efficient visual grounding transfer. Then, a standard LoRA distillation is performed over all layers to transfer multimodal reasoning capabilities. In this way, VGG-CoT decouples visual grounding from multimodal reasoning, tailoring the distillation process to the layer-wise receptiveness and limited capacity of small VLMs. Extensive experiments on several VLM architectures, including Qwen-VL and InternVL, demonstrate that VGG-CoT consistently outperforms native CoT distillation in terms of reasoning accuracy and token efficiency.

## 1 INTRODUCTION

Multimodal chain-of-thought (CoT) reasoning has emerged as a powerful paradigm for enhancing the performance of large vision-language models (VLMs) in complex tasks (Cheng et al., 2025; Zhang et al., 2025). By structuring inference as a sequence of intermediate reasoning steps, CoT enables large VLMs to generate more interpretable and accurate responses. However, transferring this capability to small VLMs through knowledge distillation remains a significant challenge. Due to their limited representational capacity, small VLMs struggle to match larger teacher models in fine-grained visual grounding and complex reasoning. Moreover, existing CoT distillation approaches (Chen et al., 2025; Wang et al., 2023) primarily rely on language-only reasoning chains without explicit visual grounding, and are therefore not readily transferable to multimodal CoT distillation for VLMs. In addition, Shao et al. (2024) provides a large-scale multimodal CoT dataset with intermediate visual grounding annotations, but does not focus on compact, layer-selective distillation approaches for small VLMs.

In this paper, we first establish a large-scale multimodal CoT dataset tailored for the VQA task, comprising 120,000 samples with length-controlled CoT chains, and region-level visual grounding annotations. Then, based on our dataset, we conduct comprehensive analyses and experiments, yielding two critical observations. Specifically, we find that VLM layers exhibit substantial variation in their receptiveness to visual grounding knowledge and that compact CoTs achieve greater reasoning efficiency and higher accuracy than lengthy native CoTs. Motivated by these observations, we propose a Visual Grounding-Guided CoT (VGG-CoT) distillation approach for VLMs, which consists of two components: (1) visual grounding receptiveness probing and (2) two-stage LoRA distillation comprising grounding distillation and reasoning distillation. First, we independently train a separate



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

LoRA adapter for each decoder layer of the student VLM for visual grounding. We then evaluate the resulting models' grounding accuracy to identify the layers most receptive to visual grounding knowledge. Next, we perform visual grounding distillation by restricting LoRA updates to the most receptive layers, thereby achieving robust grounding performance with fewer trainable parameters and preserving the pretrained visual capabilities of the remaining less receptive layers. Finally, we apply multimodal reasoning distillation across all layers to transfer the teacher VLM's multimodal chain-of-thought reasoning capability to the student VLM.

Our main contributions are summarized as follows:

- We construct a large-scale multimodal CoT dataset comprising 120,000 samples, each of which contains a compact CoT with a standardized structure, explicit length constraints, and region-level visual grounding annotations.
- We introduce VGG-CoT, a novel distillation framework that decouples visual grounding from multimodal reasoning through visual grounding receptiveness probing, visual grounding distillation, and multimodal reasoning distillation.
- We conduct extensive experiments on three VLMs across five diverse VQA task categories, demonstrating consistent improvements in both visual grounding and multimodal CoT reasoning.

## 2 Related Work

**Multimodal Chain-of-Thought.** Multimodal chain-of-thought (CoT) has emerged as an effective paradigm for enhancing the complex problem-solving capabilities of large vision-language models (VLMs) (Cheng et al., 2025; Zhang et al., 2025). Recent work relevant to multimodal reasoning spans visual reasoning benchmarks, explicit CoT supervision, visual grounding and tokenization, hallucination mitigation, multi-image reasoning, cross-modal transfer, adaptive-length CoT distillation, and related language-only long-chain distillation (Hao et al., 2025; Chen et al., 2026a; Zhang et al., 2025; Ren et al., 2025; Jiang et al., 2024; Zhang et al., 2026; Zhan et al., 2026; Peng et al., 2025; Liu et al., 2026). However, the efficient transfer of multimodal CoT capabilities to lightweight VLMs remains comparatively underexplored.

**Knowledge Distillation.** Knowledge distillation transfers knowledge from a capable teacher model to a small student model, providing an effective way to enhance the capabilities of lightweight models. In recent years, knowledge distillation has expanded beyond soft-target and sequence-level supervision to encompass generative and reasoning-aware approaches (Hinton et al., 2015; Kim & Rush, 2016; Gu et al., 2024; Chen et al., 2025; Wang et al., 2023; Agarwal et al., 2024). However, existing approaches primarily focus on language-level knowledge transfer and apply uniform adaptation across all decoder layers, without explicitly distinguishing visual grounding from subsequent multimodal reasoning.

**Visual Functions in VLMs.** Recent studies have shown that visual functions and sensitivity to hallucinations vary across decoder layers in VLMs, while others have investigated the internal mechanisms underlying CoT and the effects of reasoning supervision on multimodal reasoning (Shi et al., 2025; Dang et al., 2026; Chen et al., 2026b; Zhang et al., 2025). In this paper, we introduce layer-wise grounding receptiveness probing and integrate it with LoRA-based distillation to facilitate the transfer of multimodal CoT capabilities to lightweight models.

## 3 Dataset and Analysis

### 3.1 Dataset Establishment

In this paper, we establish the Compact Multimodal Chain-of-Thought (CM-CoT) dataset for visual question answering tasks, which consists of 120,000 samples with images, VQA annotations, and multimodal CoTs. We first collect images and VQA annotations from 12 widely used datasets spanning five task categories:

- **Visual Grounding**: Flickr30k Entities (Plummer et al., 2015) and OpenImages (Kuznetsova et al., 2020);

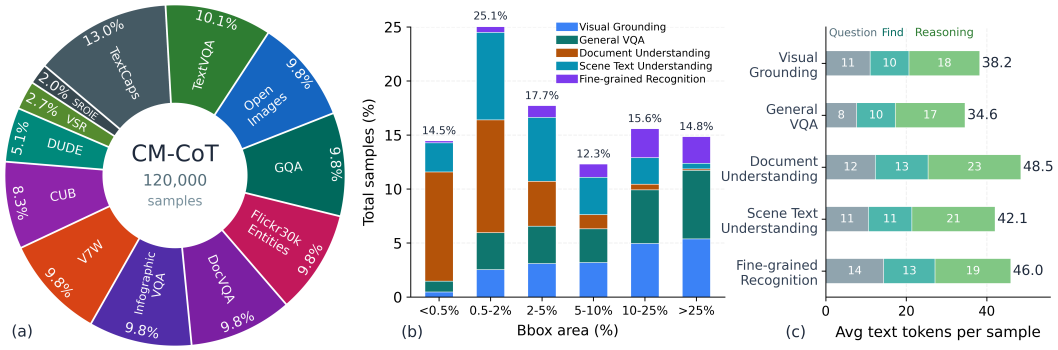



Figure 1: (a) The donut chart showing CM-CoT dataset composition by task family and source dataset. (b) The distribution of grounding bounding box areas in the CM-CoT dataset, normalized by image size. (c) The distribution of text-token composition in the CM-CoT dataset, excluding image and result tokens.

- **General VQA**: V7W (Zhu et al., 2016), GQA (Hudson & Manning, 2019), and VSR (Liu et al., 2023);
- **Document Understanding**: DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), DUDE (Landeghem et al., 2023), and SROIE (Huang et al., 2019);
- **Scene Text Understanding**: TextVQA (Singh et al., 2019) and TextCaps (Sidorov et al., 2020);
- **Fine-grained Recognition**: CUB (Wah et al., 2011).

Then, we conduct a two-stage annotation procedure to generate compact CoTs for images in our dataset. In the first stage, we use Qwen3-VL-32B-Thinking (Team, 2025) as the teacher model to produce visual annotations. Given the image, question, and answer, the teacher VLM outputs two kinds of annotations: the grounding annotation and the compact reasoning annotation. Instead of directly generating a long, free-form CoT, the teacher model is prompted to produce structured annotations that capture its evidence-selection and reasoning behavior in a concise manner. As shown in Fig. 2, the grounding annotation includes the target information, the corresponding image region, the reason why the region provides visual support for the answer, and the normalized bounding box coordinates of the selected region. In this way, the teacher VLM's evidence-selection behavior is extracted as structured and brief annotations. To further distill the teacher's multimodal CoT reasoning into a compact form suitable for small student VLMs, we prompt the teacher model to generate structured reasoning annotations that capture the task type, answer format, visual anchor, and self-correction. To this end, the original image, the question, the selected-region crop, and the grounding annotation are fed into the teacher model to generate the structured reasoning annotations. Specifically, the task type identifies the visual skill required by the question, such as attribute recognition, text reading, spatial reasoning, counting, or document field extraction. The answer format specifies the expected form of the response, such as a yes/no label, a color name, a number, a short noun phrase, or an exact text span. The visual anchor explicitly connects the reasoning to content observed in the selected region. To reduce interference from irrelevant visual content and mitigate hallucinations, we introduce a self-correction field that records a concise check against potential visual distractors, such as adjacent objects or irrelevant text lines. Finally, the resulting grounding and reasoning annotations constitute compact CoT annotations for our dataset.

As shown in Fig. 1(a), our CM-CoT dataset spans a wide range of tasks, thereby providing a comprehensive benchmark for evaluating visual reasoning capabilities. In addition, this dataset exhibits significant diversity in visual grounding granularity, with bounding box areas ranging from under 0.5% of the image area to over 25% in the full dataset. Over a third of the samples (39.55%) have bounding boxes covering less than 2% of the image area, highlighting the fine-grained nature of our dataset, as illustrated in Fig. 1(b). Moreover, the compact design of our CoT annotations ensures token-efficient reasoning, with an average length of 56.03 tokens and 98.86% of samples containing no more than 80 tokens in CoT text. This not only improves the computational efficiency of student-model inference but also facilitates the distillation of reasoning capabilities into small VLMs. For




Figure 2: Example of structured grounding and compact reasoning annotations.

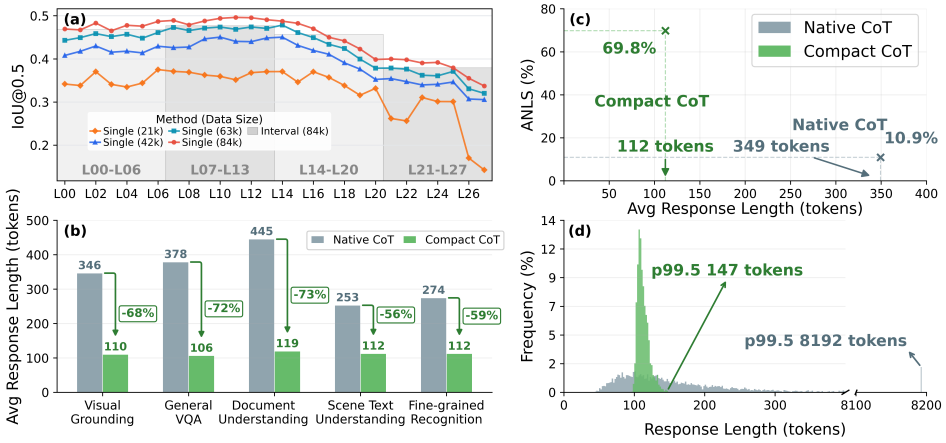



Figure 3: Comparison of grounding receptiveness and Native versus compact CoT efficiency. (a) Visual grounding performance obtained by separately probing each of the 28 decoder layers in Qwen3-VL-2B via single-layer LoRA training at each training data scale (21k, 42k, 63k, and 84k samples). (b) Average Native and compact CoT response lengths across five task categories, measured in tokens. (c) Average response length versus ANLS for Native and compact CoT. (d) Response-length distributions. All length and ANLS summaries in (b)–(d) use the same response length, measured by the number of generated text tokens.

the subsequent analyses and experiments, we partition our dataset into training, validation, and test sets at a ratio of 14:1:5 (84,000, 6,000, and 30,000 samples, respectively). We use cosine similarity between DINOv3 image embeddings to guide the split, ensuring no overlap between images in the training, validation, and test sets.

### 3.2 Data Analysis

*Finding 1:* In small VLMs, decoder layers vary substantially in their receptiveness to visual grounding knowledge during distillation, and this layer-wise receptiveness variation remains consistent across different training data scales.

*Analysis:* To investigate the layer-wise receptiveness of small VLMs to visual grounding knowledge, we conduct a probing experiment on Qwen3-VL-2B, which has 28 decoder layers. For each decoder layer, we train a lightweight LoRA adapter using our CM-CoT dataset while keeping all other decoder layers frozen. We conduct this probing experiment at four different training data scales (21k, 42k, 63k, and 84k samples) to assess the stability of layer-wise receptiveness across varying amounts of training data. Then, we evaluate the visual grounding performance of the fine-tuned models using IoU@0.5 between predicted and teacher bounding boxes. As shown in Fig. 3(a), the layer-wise receptiveness exhibits a consistent pattern across data scales: a broad plateau over lay-

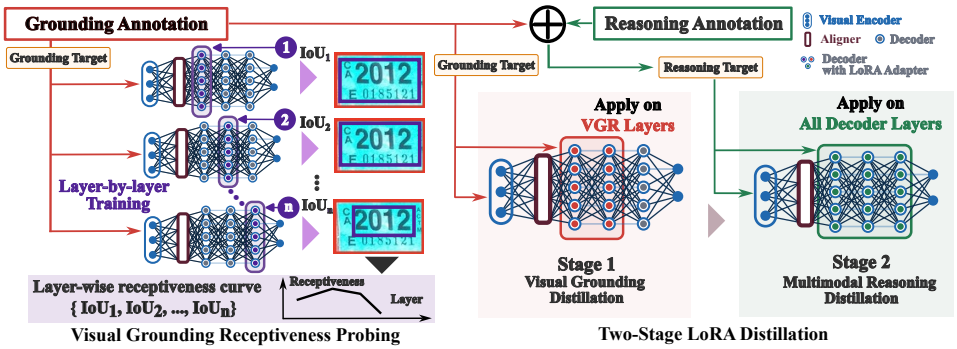



Figure 4: Overview of VGG-CoT. Receptiveness probing identifies the contiguous decoder layers most receptive to grounding knowledge. Grounding distillation updates these layers, followed by reasoning distillation across all decoder layers to generate compact CoTs and final answers.

ers 6–16, followed by a sharp decline over layers 17–27. This indicates that the lower and middle decoder layers are more receptive to visual grounding knowledge, while the upper layers are less responsive. We further perform multi-layer interval probing by partitioning the 28 decoder layers into four consecutive groups (L0–L6, L7–L13, L14–L20, and L21–L27) and training LoRA adapters for all layers within each group. The interval-level results, shown as gray bars in Fig. 3(a), closely align with the single-layer trends, indicating that the layer-wise receptiveness pattern remains consistent under multi-layer LoRA training. We also conduct a similar analysis on other VLMs with different architectures, including Qwen2.5-VL-7B and InternVL3-1B. Similar layer-wise receptiveness patterns are observed, suggesting that this pattern may generalize across VLM architectures (see the supplementary material for detailed results).

*Finding 2:* Compact CoT, with its structured and length-controlled formulation, improves both task accuracy and token efficiency in student VLMs compared with native free-form CoT.

*Analysis:* To compare the effects of native free-form CoT and our compact CoT on student model performance, we conduct experiments using Qwen3-VL-2B as the student model. Specifically, we train two instances of Qwen3-VL-2B, both with LoRA applied to all decoder layers, under identical settings except for the CoT data: one uses native CoT data and the other uses compact CoT data. After training, we evaluate both systems on the CM-CoT validation set, retaining the shortest 99.5% of responses by length for each method. Both student models receive the same original images and questions with deterministic decoding and an 8,192-token output budget. We then compute the average response length and evaluate task accuracy using Average Normalized Levenshtein Similarity (ANLS) (Biten et al., 2019) with the same normalization and answer-extraction method. As shown in Fig. 3(b), compared with native CoT training, compact CoT training significantly reduces the average response length across all five categories, with an average reduction of 67.94%, from 349.35 to 111.99 tokens. More importantly, as shown in Fig. 3(c), the model trained on compact CoT achieves significantly higher answer accuracy than its native CoT counterpart, with the ANLS score increasing from 10.87% to 69.84%. Fig. 3(d) further shows that compact CoT substantially reduces response-length variability, lowering the standard deviation from 1,170.96 to 8.08 tokens and thereby yielding more stable outputs with fewer extreme-length responses. During our experiments, we set the maximum generation length to 8,192 tokens. Under this setting, 123 of the 5,970 retained native CoT responses (2.06%) reach the limit, whereas the 99.5th percentile of compact CoT response lengths is only 147 tokens. These results demonstrate that compact CoT not only reduces token consumption but also improves answer quality, making it particularly beneficial for capacity-constrained small VLMs.

## 4 The Proposed Approach

### 4.1 Overview

In this section, we propose Visual Grounding-Guided Chain-of-Thought Distillation (VGG-CoT), a unified framework for distilling multimodal CoT reasoning capabilities into small VLMs. As

illustrated in Fig. 4, given the generated compact CoT annotations, VGG-CoT consists of two components: visual grounding receptiveness probing and two-stage LoRA distillation. First, visual grounding receptiveness probing identifies the decoder layers of the student VLM that are most receptive to visual grounding knowledge. Then, the two-stage LoRA distillation transfers the teacher's multimodal reasoning capabilities into the student model in a progressive manner: visual grounding distillation updates only the selected receptive layers to learn visual grounding, while multimodal reasoning distillation subsequently trains all decoder layers to reason based on the localized visual evidence and generate answers in the required format. In this way, VGG-CoT enables small VLMs to generate compact, visually grounded CoT responses with improved accuracy and lower token consumption on VQA tasks.

### 4.2 Visual Grounding Receptiveness Probing

Existing studies (Shi et al., 2025; Yu & Lee, 2025) have shown that different decoder layers in VLMs specialize in distinct visual functions, such as grounding and counting. Accordingly, fine-tuning layers unrelated to the target function is inefficient and may degrade existing visual capabilities in those layers. Therefore, distinguishing the layers that are most receptive to visual grounding knowledge is crucial for multimodal CoT distillation. To this end, we propose visual grounding receptiveness probing, a procedure that identifies the decoder layers that can most effectively learn visual grounding knowledge from the teacher VLM. Given a student model, we independently attach a lightweight LoRA adapter to each decoder layer while keeping all other parameters frozen. Each adapter is trained with the grounding annotations in CM-CoT, and all layers use the same training split, optimization configuration, and LoRA rank. After training, we evaluate each layer-specific adapter on the validation set by comparing its predicted boxes with the teacher reference boxes. The resulting grounding accuracy is used as the grounding receptiveness of the corresponding layer. We repeat this process across all decoder layers to obtain a layer-wise grounding receptiveness curve, which characterizes how effectively individual layers acquire and generalize visual grounding knowledge.

We therefore partition the curve into consecutive intervals, allowing the selected layers to capture a coherent response pattern across neighboring layers. We use the Bayesian information criterion (BIC) (Schwarz, 1978; Yao, 1988) to automatically determine the number and boundaries of these intervals while penalizing unnecessary partitions. The layers in the interval with the highest mean validation receptiveness are then designated as the Visual Grounding Receptiveness (VGR) Layers. For Qwen3-VL-2B, BIC produces five intervals, $L_0$–$L_5$, $L_6$–$L_{16}$, $L_{17}$–$L_{19}$, $L_{20}$–$L_{24}$, and $L_{25}$–$L_{27}$, whose mean IoU@0.5 values are 0.4731, 0.4872, 0.4382, 0.3957, and 0.3572, respectively. The second interval therefore achieves the highest grounding performance, and its constituent layers are selected as the VGR Layers. We can apply the same procedure to the other student architectures to identify their respective contiguous VGR Layers.

### 4.3 Two-Stage LoRA Distillation

To transfer multimodal reasoning capabilities to small VLMs, we extend the rationale-based distillation framework of Hsieh et al. (2023) from text-only reasoning to multimodal CoT supervision. Rather than distilling the teacher response as a single target, we decompose the learning process into visual grounding and subsequent visual evidence-based reasoning. The student VLM is trained offline on pre-generated teacher annotations, without requiring teacher logits or intermediate representations. Specifically, our two-stage LoRA distillation comprises grounding distillation and reasoning distillation. The first stage updates only the contiguous VGR Layers selected by visual grounding receptiveness probing, whereas the reasoning stage updates all decoder layers. This design confines grounding adaptation to the VGR Layers to avoid disrupting non-receptive layers, while enabling broader adaptation across all decoder layers for multimodal reasoning and answer generation.

**Visual Grounding Distillation.** The grounding stage is designed to teach the student VLM to identify the visual evidence required by the question. During this stage, LoRA adapters are attached only to the contiguous decoder-layer interval selected by visual grounding receptiveness probing. The visual encoder, the aligner, and all remaining decoder layers are kept frozen. Given the original image and the question, the model is trained with response-level causal language modeling, where the grounding annotation serves as the assistant-side target. Specifically, the grounding annotations

Table 1: Visual grounding performance on six representative datasets. Avg. is the arithmetic mean over the six datasets. The best and second-best values are bolded and underlined, respectively. All results are reported as percentages (%).

| Model | Approach | OpenImages | | V7W | | Flickr30k | | TextCaps | | CUB | | TextVQA | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $IoU_{.3}$ | $IoU_{.5}$ | $IoU_{.3}$ | $IoU_{.5}$ | $IoU_{.3}$ | $IoU_{.5}$ | $IoU_{.3}$ | $IoU_{.5}$ | $IoU_{.3}$ | $IoU_{.5}$ | $IoU_{.3}$ | $IoU_{.5}$ | $IoU_{.3}$ | $IoU_{.5}$ |
| Qwen2.5-VL 7B | Token Swapping | 46.58 | 33.91 | 39.04 | 28.04 | 41.98 | 26.55 | 44.04 | 21.23 | **71.21** | 64.05 | 46.14 | 23.38 | 48.16 | 32.86 |
| | Few-Shot CoT | 37.21 | 32.83 | 35.42 | 29.78 | 36.54 | 29.06 | 28.08 | 17.92 | 70.17 | **69.57** | 29.04 | 18.41 | 39.41 | 32.93 |
| | Vanilla Layers | 47.32 | 34.48 | 40.13 | 30.05 | 42.36 | 27.89 | 45.04 | 21.97 | 66.01 | 56.42 | 47.03 | 22.32 | 47.98 | 32.19 |
| | **VGG-CoT** | **51.23** | **40.14** | **42.55** | **33.30** | **44.94** | **31.78** | **52.58** | **29.92** | 67.57 | 62.14 | **54.65** | **32.26** | **52.25** | **38.26** |
| Qwen3-VL 2B | Token Swapping | 39.13 | 31.28 | 42.01 | 32.89 | **49.07** | 35.85 | 60.24 | 43.62 | 48.90 | 36.55 | 63.87 | 47.65 | 50.54 | 37.97 |
| | Few-Shot CoT | 41.15 | 34.78 | 41.97 | 34.12 | 48.31 | 35.71 | 51.52 | 38.66 | **91.60** | **87.49** | 55.85 | 42.68 | 55.07 | 45.57 |
| | Vanilla Layers | **53.08** | **44.46** | 44.67 | 36.78 | 48.76 | **37.05** | 62.98 | 47.55 | 65.69 | 59.54 | 66.95 | 51.46 | 57.02 | 46.14 |
| | **VGG-CoT** | 50.66 | 42.67 | **45.36** | **37.64** | 46.42 | 33.99 | **67.80** | **56.77** | 73.33 | 69.77 | **73.16** | **62.46** | **59.45** | **50.55** |
| InternVL3 1B | Token Swapping | 39.70 | 24.74 | 35.21 | 24.35 | 40.74 | 22.45 | 12.74 | 4.03 | 62.69 | 47.58 | 12.72 | 3.94 | 33.97 | 21.18 |
| | Few-Shot CoT | 7.85 | 2.09 | 7.34 | 2.66 | 11.98 | 3.37 | 3.19 | 0.64 | 2.16 | 0.08 | 3.50 | 0.69 | 6.00 | 1.59 |
| | Vanilla Layers | 45.77 | 34.11 | 38.66 | 29.00 | 42.39 | 28.68 | 21.60 | 8.17 | 62.18 | 52.50 | 24.07 | 9.43 | 39.11 | 26.98 |
| | **VGG-CoT** | **49.92** | **38.25** | **41.12** | **32.07** | **43.56** | **30.89** | **29.26** | **12.49** | **71.49** | **65.05** | **31.78** | **13.51** | **44.52** | **32.04** |

explicitly supervise the identification and localization of answer-relevant visual content. Restricting LoRA updates to the selected interval concentrates this supervision on the layers identified as most receptive to grounding knowledge.

**Multimodal Reasoning Distillation.** Reasoning distillation follows grounding distillation and enables the student to reason effectively over grounded visual evidence. Starting from the grounding-distilled checkpoint, we attach LoRA adapters to all decoder layers and optimize the model with the compact CoT reasoning annotations. Specifically, each training sample contains the original image and question, together with the grounding crop and its corresponding annotation as additional context for reasoning distillation. The corresponding reasoning annotation, in turn, serves as the target for response-level causal language modeling. During reasoning distillation, the LoRA adapters attached to all decoder layers are updated, while the visual encoder and aligner remain frozen.

## 5 Experimental Results

In this section, we conduct experiments to evaluate VGG-CoT with Qwen3-VL-2B (Team, 2025), Qwen2.5-VL-7B (Bai et al., 2025), and InternVL3-1B (Zhu et al., 2025) on the test set of our CM-CoT dataset. At test time, the bounding box and CoT for both stage 1 and stage 2 are generated by the student model. For visual grounding, we use IoU@0.3 and IoU@0.5 between predicted and annotated regions as metrics. For VQA, we adopt ANLS (Biten et al., 2019) and token-level F1 (Rajpurkar et al., 2016) as metrics. We compare VGG-CoT with Native SeqKD (Kim & Rush, 2016), Few-Shot CoT (Wei et al., 2022), and Token Swapping (Shi et al., 2025). We further compare VGG-CoT with two additional baselines: Vanilla Layers, which applies grounding distillation to the last decoder layers with the same number of trainable parameters as the VGR Layers, and Original Model, which uses the original student model without distillation. During training, all the distillation approaches use CM-CoT data except Native SeqKD, which uses Native CoT responses. All approaches are evaluated under the same held-out evaluation protocol and answer-scoring rules.

### 5.1 Evaluation on Grounding Tasks

We evaluate the grounding performance of VGG-CoT against the baselines across three VLMs on all twelve sub-datasets of the CM-CoT test set. Table 1 reports the results on six representative datasets. OpenImages, V7W, Flickr30k Entities, GQA, and CUB use source-dataset boxes as IoU references, while the remaining datasets use teacher-generated boxes since they do not provide task-aligned reference boxes. Note that the average results are computed over all six datasets. VGG-CoT achieves the highest average score across the six datasets for every model under both IoU thresholds. On average, VGG-CoT improves over Token Swapping by 12.58 and 5.40 percentage points at IoU@0.5 for Qwen3-VL-2B and Qwen2.5-VL-7B, respectively. By contrast, Vanilla Layers achieves lower average performance across the six datasets than VGG-CoT for all three models, verifying that the non-receptive decoder layers are not suitable for grounding distillation. The above results demonstrate that VGG-CoT effectively transfers visual grounding knowledge from the teacher to the stu-

Table 2: Reasoning performance on six representative datasets. Avg. is the arithmetic mean over the six datasets. The best and second-best values are bolded and underlined, respectively. All results are reported as percentages (%).

| Model | Approach | OpenImages | | V7W | | Flickr30k | | TextCaps | | CUB | | TextVQA | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ANLS | F1 | ANLS | F1 | ANLS | F1 | ANLS | F1 | ANLS | F1 | ANLS | F1 | ANLS | F1 |
| Qwen2.5-VL 7B | Original Model | 25.50 | 21.49 | 47.23 | 44.39 | 37.13 | 40.94 | 76.11 | 76.88 | 74.66 | 74.65 | **87.90** | **86.18** | 58.09 | 57.42 |
| | Native SeqKD | 5.73 | 5.41 | 10.44 | 14.06 | 20.30 | 26.74 | 8.88 | 13.43 | 0.67 | 2.12 | 12.48 | 17.50 | 9.75 | 13.21 |
| | Token Swapping | 53.37 | 43.41 | 51.27 | 46.91 | 42.12 | 44.40 | 76.73 | 76.46 | 75.77 | 75.77 | 85.46 | 83.49 | 64.12 | 61.74 |
| | Few-Shot CoT | 20.92 | 16.52 | 34.19 | 33.31 | 35.40 | 37.02 | 70.09 | 71.18 | 32.54 | 37.22 | 75.00 | 75.29 | 44.69 | 45.09 |
| | Vanilla Layers | 66.88 | 57.64 | 52.65 | 47.83 | 44.64 | 46.83 | 76.89 | 76.75 | 79.17 | 79.17 | 84.32 | 82.41 | 67.43 | 65.11 |
| | **VGG-CoT** | **68.50** | **58.97** | **52.95** | **48.07** | **45.19** | **47.39** | **77.08** | **77.20** | **80.25** | **80.25** | 84.91 | 82.90 | **68.15** | **65.80** |
| Qwen3-VL 2B | Original Model | 50.11 | 38.32 | **53.05** | **45.28** | 22.82 | 23.44 | **76.16** | **75.91** | 77.03 | 76.99 | **85.83** | **82.85** | 60.83 | 57.13 |
| | Native SeqKD | 6.98 | 6.53 | 11.87 | 15.45 | 22.15 | 28.62 | 9.52 | 14.09 | 0.69 | 2.21 | 12.47 | 17.71 | 10.61 | 14.10 |
| | Token Swapping | 41.72 | 31.60 | 48.65 | 42.08 | 29.53 | 30.60 | 71.85 | 71.78 | 75.09 | 75.14 | 78.60 | 76.39 | 57.57 | 54.60 |
| | Few-Shot CoT | 24.16 | 19.51 | 40.67 | 37.58 | 33.62 | 34.38 | 70.06 | 70.01 | 9.30 | 15.31 | 75.01 | 74.24 | 42.14 | 41.84 |
| | Vanilla Layers | 65.18 | 54.76 | 50.79 | 44.47 | 41.52 | 44.04 | 72.58 | 71.62 | 79.57 | 79.57 | 81.66 | 78.25 | 65.22 | 62.12 |
| | **VGG-CoT** | **67.76** | **57.99** | 51.63 | 45.08 | **41.82** | **44.26** | 74.56 | 73.63 | **80.24** | **80.24** | 83.15 | 79.82 | **66.53** | **63.50** |
| InternVL3 1B | Original Model | 31.20 | 23.13 | 45.36 | 42.32 | 38.05 | 40.68 | 71.63 | **72.42** | 71.99 | 72.05 | **84.86** | **84.92** | 57.18 | 55.92 |
| | Native SeqKD | 7.91 | 7.40 | 12.98 | 16.29 | 23.66 | 29.40 | 9.99 | 14.38 | 0.73 | 2.28 | 13.38 | 18.80 | 11.44 | 14.76 |
| | Token Swapping | 53.24 | 41.98 | 44.82 | 36.82 | 19.95 | 20.91 | 66.87 | 65.50 | 64.07 | 64.14 | 75.29 | 73.56 | 54.04 | 50.48 |
| | Few-Shot CoT | 17.92 | 13.62 | 22.17 | 19.34 | 24.83 | 26.05 | 38.61 | 40.11 | 14.54 | 11.17 | 40.77 | 41.74 | 26.47 | 25.34 |
| | Vanilla Layers | 65.00 | 56.30 | 49.43 | 42.44 | 42.86 | 45.27 | 71.11 | 70.07 | 71.45 | 71.45 | 80.55 | 79.30 | 63.40 | 60.81 |
| | **VGG-CoT** | **70.82** | **63.01** | **50.51** | **43.53** | **44.75** | **47.56** | **71.91** | 71.29 | **75.89** | **75.89** | 81.42 | 80.31 | **65.88** | **63.60** |

dent model, enabling the student to accurately identify the visual regions required for answering questions. In addition, the superior average performance of VGG-CoT across six datasets and three VLMs demonstrates its generalizability across datasets and model architectures.

### 5.2 Evaluation on Reasoning Tasks

We evaluate the reasoning performance of VGG-CoT against the baselines across three VLMs on six sub-datasets of the CM-CoT test set. For quantitative comparison, we adopt Average Normalized Levenshtein Similarity (ANLS) and token-level F1 as evaluation metrics. The results are reported in Table 2. VGG-CoT achieves the highest average performance for every model under both reasoning metrics. VGG-CoT improves average ANLS and F1 over Vanilla Layers by 0.72–2.48 and 0.69–2.79 points, respectively, while also consistently outperforming Token Swapping. Native SeqKD performs substantially worse because its verbose, unconstrained CoTs may obscure or truncate the final answer, as illustrated in Fig. 3(c,d). Although the original models remain superior on TextVQA for all three student architectures, VGG-CoT achieves the best overall performance in most settings. These results demonstrate that VGG-CoT effectively transfers multimodal reasoning capabilities from the teacher to student models and generalizes well across datasets and model architectures.

Figure 5 shows an example of distillation results for a store-section question. The base model distilled with Native CoT produces an incorrect answer, while VGG-CoT uses the grounding crop and obtains the correct answer. As is shown, the student using Native SeqKD initially recognizes the "Baby" section sign in the image correctly, but is then misled by the prominent "Rollback" promotion and selects it instead of the section sign and invents "New" and "Clothes" signs absent from the image. By contrast, our VGG-CoT exposes this contradiction during self-correction, rejects the promotional distraction, and recovers the correct answer. This example demonstrates that VGG-CoT enhances multimodal CoT reasoning by enabling small VLMs not only to identify relevant visual evidence but also to explicitly expose and correct distractor-driven errors during inference.

### 5.3 Ablation Study

We conduct ablation studies to evaluate the effectiveness of the individual components of VGG-CoT. Specifically, we perform all ablations on Qwen3-VL-2B and evaluate the resulting variants across all twelve sub-datasets of the CM-CoT test set. The arithmetic mean over the twelve datasets is reported in Table 3. As shown, directly applying the native CoT supervision to train the student model achieves only 12.44 ANLS, indicating that the lengthy native CoT is not sufficient for reasoning

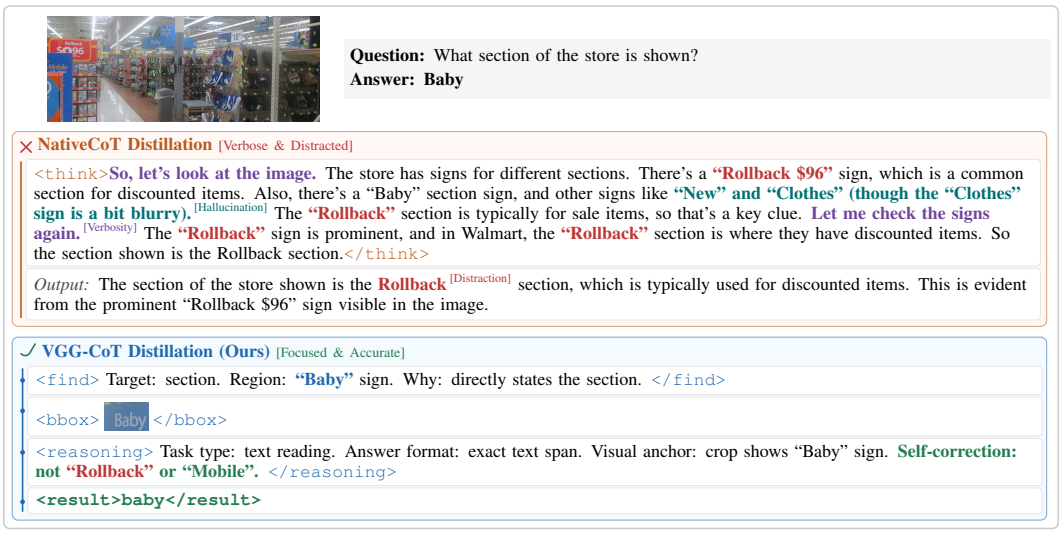



Figure 5: A qualitative example showing how VGG-CoT identifies and corrects an error.

Table 3: Ablation results of our VGG-CoT approach.

| CoT | Two-stage | Layers | | | IoU@0.3 | IoU@0.5 | ANLS | F1 |
|---|---|---|---|---|---|---|---|---|
| | | Base $L17$–$L27$ | Middle $L9$–$L19$ | VGR $L6$–$L16$ | | | | |
| — | — | — | — | — | – | – | 65.48 | 60.12 |
| ✓ | — | — | — | — | – | – | 12.44 | 15.33 |
| ✓ | ✓ | ✓ | — | — | 49.45 | 37.40 | 65.88 | 61.07 |
| ✓ | ✓ | — | ✓ | — | 53.67 | 42.28 | 66.87 | 62.21 |
| ✓ | ✓ | — | — | ✓ | **54.00** | **42.62** | **67.20** | **62.62** |

transfer. This may be because the small student VLM struggles to learn the teacher's complex reasoning patterns. In the third row, we replace the native CoT with our compact CoT and apply our two-stage distillation framework, which improves ANLS to 65.88 and IoU@0.5 to 37.40. Note that this is achieved by training a fixed range of decoder layers (layers 17–27, denoted as Base) in the grounding distillation stage. We also compare the effects of training the middle decoder-layer interval (layers 9–19, denoted as Middle) and the VGG-selected interval (layers 6–16, denoted as VGR), with the latter identified through visual grounding receptiveness probing. We can observe that the Middle variant improves ANLS to 66.87, while the VGR variant further improves ANLS to 67.20 and IoU@0.5 to 42.62. The results demonstrate that all the components effectively improve the reasoning and grounding capabilities of the student VLM.

## 6 Conclusion

In this paper, we presented VGG-CoT, a visual grounding-guided chain-of-thought distillation framework for transferring multimodal reasoning capabilities to small VLMs. Specifically, we constructed a large-scale multimodal CoT dataset comprising 120,000 samples, each containing a compact CoT with a standardized structure and explicit length constraints, along with region-level visual grounding annotations. We then conducted a thorough analysis of our dataset and found that visual grounding receptiveness varies substantially across decoder layers and that compact CoTs achieve greater reasoning efficiency and higher accuracy than lengthy native CoTs. Motivated by these findings, VGG-CoT first probes visual grounding receptiveness across the student model's decoder layers and then performs two-stage LoRA distillation to transfer multimodal reasoning capabilities from the teacher to the student. Extensive experiments on Qwen and InternVL models demonstrate that VGG-CoT effectively transfers multimodal CoT reasoning to small VLMs, achieving stronger overall visual grounding and reasoning performance than the baseline approaches.

## AI USE STATEMENT

Generative AI tools were used for text polishing, dataset annotation, and CoT generation. The authors reviewed the AI-assisted content and take responsibility for the final paper and dataset.

## ETHICS STATEMENT

This study uses publicly available datasets and does not involve human-subject experiments. Source images are not redistributed and remain subject to their original datasets' terms of use.

## REPRODUCIBILITY STATEMENT

The appendix describes the dataset splits, training configurations, evaluation metrics, and layer-selection procedure. It also describes the accompanying supplementary code and sample annotations for reproducing the reported analyses.

## REFERENCES


Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. *ICLR*, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan 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. *CoRR*, abs/2502.13923, 2025. doi: 10.48550/ARXIV.2502.13923. URL https://doi.org/10.48550/arXiv.2502.13923.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marcal Rusinol, Minesh Mathew, C.V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1563–1570, 2019. doi: 10.1109/ICDAR.2019.00251.

Liang Chen, Weichu Xie, Yiyan Liang, Hongfeng He, Hans Zhao, Zhibo Yang, Zhiqi Huang, Haoning Wu, Haoyu Lu, Y. Charles, Yiping Bao, Yuantao Fan, Guopeng Li, Haiyang Shen, Xuanzhong Chen, Wendong Xu, Shuzheng Si, Zefan Cai, Wenhao Chai, Ziqi Huang, Fangfu Liu, Tianyu Liu, Baobao Chang, Xiaobo Hu, Kaiyuan Chen, Yixin Ren, Yang Liu, Yuan Gong, and Kuan Li. Babyvision: Visual reasoning beyond language. *CoRR*, abs/2601.06521, 2026a. doi: 10.48550/ARXIV.2601.06521. URL https://doi.org/10.48550/arXiv.2601.06521.

Xi Chen, Aske Plaat, and Niki Van Stein. How does chain of thought think? Mechanistic interpretability of chain-of-thought reasoning with sparse autoencoding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30297–30305, 2026b. doi: 10.1609/aaai.v40i36.40281.

Xinghao Chen, Zhijing Sun, Wenjin Guo, Miaoran Zhang, Yanjun Chen, Yirong Sun, Hui Su, Yijie Pan, Dietrich Klakow, Wenjie Li, and Xiaoyu Shen. Unveiling the key factors for distilling chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15094–15119, 2025. doi: 10.18653/v1/2025.findings-acl.782.

Kanzhi Cheng, YanTao Li, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. Vision-language models can self-improve reasoning via reflection. 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)*, pp. 8876–8892, 2025. doi: 10.18653/v1/2025.naacl-long.447.

TianTian Dang, Chao Bi, Shufan Shen, Jinzhe Liu, Qingming Huang, and Shuhui Wang. Locate-then-sparsify: Attribution guided sparse strategy for visual hallucination mitigation. *CoRR*, abs/2603.16284, 2026. doi: 10.48550/ARXIV.2603.16284.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. *ICLR*, 2024.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can MLLMs reason in multimodality? EMMA: An enhanced MultiModal ReAsoning benchmark. *ICML*, 2025.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

Cheng-Yu Hsieh, Chun-Liang Li, Chih kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023. doi: 10.18653/v1/2023.findings-acl.507.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520, 2019. doi: 10.1109/icdar.2019.00244.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702, 2019. doi: 10.1109/cvpr.2019.00686.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27026–27036, 2024. doi: 10.1109/cvpr52733.2024.02553.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016. doi: 10.18653/v1/d16-1139.

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. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. doi: 10.1007/s11263-020-01316-z.

Jordy Van Landeghem, Rafał Powalski, Rubèn Tito, Dawid Jurkiewicz, Matthew Blaschko, Łukasz Borchmann, Mickaël Coustaty, Sien Moens, Michał Pietruszka, Bertrand Ackaert, Tomasz Stanisławek, Paweł Józiak, and Ernest Valveny. Document understanding dataset and evaluation (dude). In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19471–19483, 2023. doi: 10.1109/iccv51070.2023.01789.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. doi: 10.1162/tacl_a_00566.

Zhenghao Liu, Zhuoyang Wu, Xinze Li, Yukun Yan, Shuo Wang, Zulong Chen, Yu Gu, Ge Yu, and Maosong Sun. Long-chain reasoning distillation via adaptive prefix alignment. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18041–18059, 2026. doi: 10.18653/v1/2026.acl-long.822.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208, 2021. doi: 10.1109/wacv48630.2021.00225.

Minesh Mathew, Viraj Bagal, Ruben Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. Infographicvqa. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2582–2591, 2022. doi: 10.1109/wacv51458.2022.00264.

Yi Peng, Chris, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, Li Ge, Rongxian Zhuang, Xuchen Song, Yang Liu, and Yahui Zhou. Skywork R1V: pioneering multimodal reasoning with chain-of-thought. *CoRR*, abs/2504.05599, 2025. doi: 10.48550/ARXIV.2504.05599. URL https://doi.org/10.48550/arXiv.2504.05599.

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 *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2641–2649, 2015. doi: 10.1109/iccv.2015.303.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016. doi: 10.18653/v1/D16-1264.

Xiangxuan Ren, Zhongdao Wang, Liping Hou, Pin Tang, Guoqing Wang, and Chao Ma. Grounding everything in tokens for multimodal large language models. *CoRR*, 2025. doi: 10.48550/ARXIV.2512.10554.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. doi: 10.1214/aos/1176344136.

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 *Advances in Neural Information Processing Systems*, 2024.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision function layer in multimodal LLMs. *CoRR*, abs/2509.24791, 2025. doi: 10.48550/ARXIV.2509.24791.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: A dataset for image captioning with reading comprehension. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347 of *Lecture Notes in Computer Science*, pp. 742–758. Springer, 2020. doi: 10.1007/978-3-030-58536-5_44. URL https://doi.org/10.1007/978-3-030-58536-5_44.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318, 2019. doi: 10.1109/cvpr.2019.00851.

Qwen Team. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025. doi: 10.48550/ARXIV.2511.21631. URL https://doi.org/10.48550/arXiv.2511.21631.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. Scott: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5546–5558, 2023. doi: 10.18653/v1/2023.acl-long.304.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800.

Yi-Ching Yao. Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, 6(3):181–189, 1988. doi: 10.1016/0167-7152(88)90118-6.

Zhuoran Yu and Yong Jae Lee. How multimodal llms solve image tasks: A lens on visual grounding, task reasoning, and answer decoding. *CoRR*, abs/2508.20279, 2025. doi: 10.48550/ARXIV.2508.20279. URL https://doi.org/10.48550/arXiv.2508.20279.

Yu-Liang Zhan, Xinyu Tang, Han Wan, Jian Li, Jirong Wen, and Hao Sun. L2v-cot: Cross-modal transfer of chain-of-thought reasoning via latent intervention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(15):12358–12366, 2026. doi: 10.1609/aaai.v40i15.38228.

Guanghao Zhang, Tao Zhong, Yan Xia, Mushui Liu, Zhelun Yu, Haoyuan Li, Wanggui He, Dong She, Yi Wang, and Hao Jiang. Cmmcot: Enhancing complex multi-image comprehension via multi-modal chain-of-thought and memory augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(15):12430–12438, 2026. doi: 10.1609/aaai.v40i15.38236.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1631–1662, 2025. doi: 10.18653/v1/2025.acl-long.82.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4995–5004, 2016. doi: 10.1109/cvpr.2016.540.

# A Dataset Details

This supplementary material provides additional details and results that complement the main paper. Sec. A provides dataset details, including the dataset composition, annotation structure, student training, and experimental environment. Sec. B presents technical details of the proposed Visual Grounding Receptiveness Probing and Visual Grounding Receptiveness Curves. Sec. C presents more experimental results across all twelve source datasets. Sec. D describes the external code and sample-data package used for reproducibility. Finally, Sec. E provides several case studies.

**Dataset Composition.** Our CM-CoT benchmark comprises 120,000 samples collected from 12 widely used datasets covering five task categories: visual grounding (Flickr30k Entities for region-to-phrase grounding and OpenImages for box localization), general VQA (GQA for scene-graph compositional reasoning, V7W for pointed and grounded visual question answering, and VSR for spatial relation discrimination), document understanding (DocVQA for visual question answering on document images, InfoVQA for structural reasoning over infographic images, DUDE for document understanding in complex layouts, and SROIE for key information extraction from receipts), scene text understanding (TextVQA for reading and reasoning on scene text and TextCaps for captioning grounded on scene text), and fine-grained recognition (CUB for species and attribute recognition). Table 4 reports the number of samples drawn from each source dataset. Accordingly, our dataset is split into 84,000 training, 6,000 validation, and 30,000 test samples, following a 14:1:5 ratio.

For grounding results in the main and ablation tables, Flickr30k Entities, V7W, CUB, GQA, and OpenImages use available source-dataset region or object boxes; VSR, TextVQA, TextCaps, DocVQA, DUDE, SROIE, and InfoVQA use our teacher-generated boxes.

Table 4: Sample distribution across source datasets in our CM-CoT benchmark.

| Task Category | Source Dataset | # Samples |
|---|---|---|
| Visual Grounding | Flickr30k Entities | 11,756 |
| | OpenImages | 11,768 |
| General VQA | GQA | 11,758 |
| | V7W | 11,754 |
| | VSR | 3,269 |
| Document Understanding | DocVQA | 11,755 |
| | InfoVQA | 11,755 |
| | DUDE | 6,168 |
| | SROIE | 2,345 |
| Scene Text Understanding | TextVQA | 12,157 |
| | TextCaps | 15,600 |
| Fine-grained Recognition | CUB | 9,915 |
| | Total | 120,000 |

**Annotation Structure.** Each example contains an image, a question, a grounding annotation, and a Compact Chain-of-Thought (CoT) annotation. The grounding annotation records the target, relevant region, rationale, and normalized bounding box. The compact CoT annotation provides the task type, answer format, visual anchor, and self-correction, followed by the final result.

**Student Training.** We evaluate VGG-CoT across three student models: Qwen3-VL-2B (Team, 2025), Qwen2.5-VL-7B (Bai et al., 2025), and InternVL3-1B (Zhu et al., 2025), all trained with LoRA adapters following the two-stage distillation procedure described in the main text. The shared training configuration and the stage-specific settings are reported in Tables 5 and 6. As summarized in Table 5, the frozen modules are the visual encoder and the aligner, and the LoRA target modules are the attention and feed-forward projections (q, k, v, o; gate, up, down); the grounding stage attaches LoRA adapters only to the selected Visual Grounding Receptiveness (VGR) Layers, while the reasoning stage attaches LoRA adapters to all decoder layers. The effective batch size is the product of the per-device batch size, the gradient accumulation steps, and the number of distributed

Table 5: Detailed training configuration. Qwen3-VL-2B, InternVL3-1B and Qwen2.5-VL-7B are trained using CANN on Ascend NPUs.

| Parameter | Value |
|---|---|
| LoRA rank | 8 |
| LoRA alpha | 32 |
| LoRA dropout | 0.05 |
| Numerical precision | bfloat16 |
| Frozen modules | Visual encoder; Aligner |
| Target modules | q, k, v, o; gate, up, down |
| Random seed | 42 |
| Optimizer | AdamW |
| DeepSpeed strategy | ZeRO-2 |

Table 6: Stage-specific training configurations for the three student models.

| Model | Stage | Effective batch | LR ($\times 10^{-4}$) | Epochs | Max seq. length |
|---|---|---|---|---|---|
| Qwen3-VL-2B | Grounding | 32 | 1.0 | 1 | 2048 |
| | Reasoning | | 1.0 | 1 | 8192 |
| Qwen2.5-VL-7B | Grounding | 32 | 1.0 | 1 | 2048 |
| | Reasoning | | 1.0 | 1 | 8192 |
| InternVL3-1B | Grounding | 32 | 1.0 | 1 | 2048 |
| | Reasoning | | 1.0 | 1 | 8192 |

workers. Training in both stages uses the CM-CoT training set, with each input image represented by at most 256 visual tokens.

**Experimental Environment.** Experiments are conducted on eight Ascend 910B3 NPUs through the CANN execution environment. The operating system is Ubuntu 22.04.4 LTS (Linux 5.15). The software stack consists of Python 3.10.19, PyTorch 2.9.0, Transformers 4.57.6, PEFT 0.18.0, and DeepSpeed 0.18.2. All models are trained with bfloat16 precision.

## B Visual Grounding Receptiveness Probing

We conduct the same layer-wise visual grounding receptiveness probing on the three reported student architectures, Qwen3-VL-2B, Qwen2.5-VL-7B, and InternVL3-1B, and report the results in Figure 6. The two Qwen sweeps cover 28 decoder layers ($L_0$–$L_{27}$), whereas the InternVL3-1B sweep covers 24 layers ($L_0$–$L_{23}$). As shown, the other two architectures exhibit a trend similar to that of Qwen3-VL-2B, consistent with *Finding 1* in our main text and suggesting that this layer-wise receptiveness pattern generalizes across VLM architectures.

**Identification Approach.** To identify decoder layers that are most receptive to visual grounding knowledge during distillation, we develop a novel identification approach. As mentioned in the main text, we partition the layer-wise grounding receptiveness curve into consecutive intervals, allowing the selected layers to capture a coherent response pattern across neighboring layers. Specifically, we use the Bayesian information criterion (BIC) (Schwarz, 1978; Yao, 1988) to automatically determine the number and boundaries of these intervals while penalizing unnecessary partitions. To avoid extremely small trainable parameter budgets, we constrain each interval to contain at least three consecutive layers. For a candidate partition $\mathcal{P} = \{I_1, \ldots, I_K\}$, the average receptiveness of interval $I_j$ and the within-interval sum of squared errors (SSE) are defined as

$$\hat{\mu}_j = \frac{1}{|I_j|} \sum_{\ell \in I_j} r_\ell, \mathrm{SSE}(\mathcal{P}) \quad = \sum_{j=1}^{K} \sum_{\ell \in I_j} (r_\ell - \hat{\mu}_j)^2, \tag{1}$$

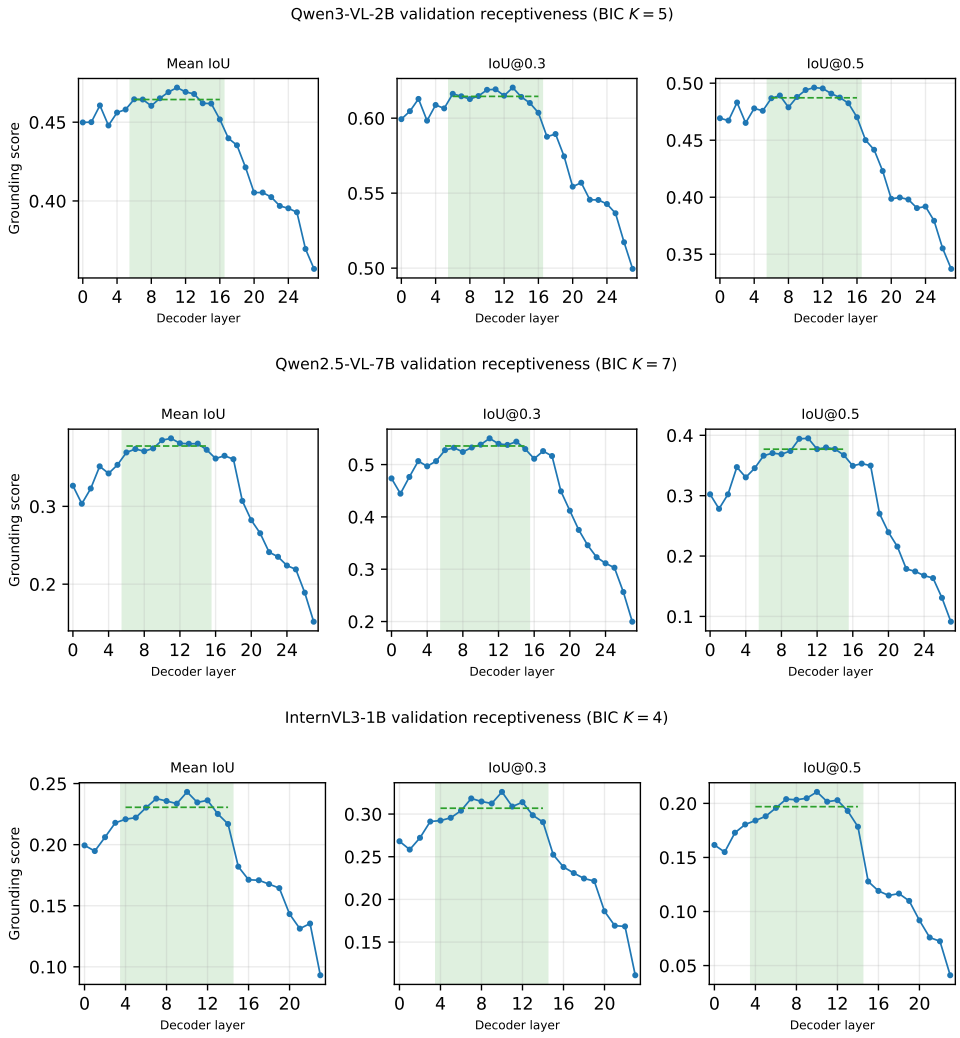



Figure 6: Visual Grounding Receptiveness Curves for Qwen3-VL-2B, Qwen2.5-VL-7B, and InternVL3-1B. The green band is the highest-mean BIC interval, selected as the VGR Layers for Stage 1.

where $r_\ell$ is the validation IoU@0.5 of layer $\ell$. Under the Gaussian step-model assumption, the corresponding criterion is

$$\begin{aligned} \mathrm{BIC}(\mathcal{P}) &= n \ln(\mathrm{SSE}(\mathcal{P})/n) + k \ln(n), \\ n &= L, k = 2K - 1. \end{aligned} \tag{2}$$

Here, $n = L$ denotes the number of decoder layers: $L = 28$ for Qwen3-VL-2B and Qwen2.5-VL-7B, and $L = 24$ for InternVL3-1B. The $k = 2K - 1$ parameters comprise $K$ interval means and $K - 1$ boundaries. For Qwen3-VL-2B, the best candidates with $K = 1, 2, 3, 4, 5$ have SSE values $0.061524$, $0.009311$, $0.005603$, $0.002965$, and $0.002192$, yielding BIC values $-168.043$, $-214.248$, $-221.808$, $-232.958$, and $-234.760$. The remaining feasible candidates ($K = 6$–$9$) have BIC values $-230.975$, $-226.393$, $-219.730$, and $-209.541$, so $K = 5$ is selected. Its partition is $\mathcal{P}_5 = \{L_0\text{–}L_5, L_6\text{–}L_{16}, L_{17}\text{–}L_{19}, L_{20}\text{–}L_{24}, L_{25}\text{–}L_{27}\}$, with respective mean IoU@0.5 values $0.4731$, $0.4872$, $0.4382$, $0.3957$, and $0.3572$. For this partition, $k = 2K - 1 = 9$, and $\mathrm{BIC}(\mathcal{P}_5) = 28 \ln(0.002191541/28) + 9 \ln(28) \approx -234.760$. Accordingly, we select the interval with the highest mean receptiveness, $L_6$–$L_{16}$, as the Visual Grounding Receptiveness (VGR) Layers. Similarly, applying this procedure to the other student architectures yields $L_6$–$L_{15}$ for Qwen2.5-VL-7B and $L_4$–$L_{14}$ for InternVL3-1B.

## C More Experimental Results

**Comparison Approaches.** We compare VGG-CoT with Native SeqKD (Kim & Rush, 2016), Few-Shot CoT (Wei et al., 2022), and Token Swapping (Shi et al., 2025). We further compare VGG-CoT with two additional baselines: Vanilla Layers, which applies grounding distillation to the last decoder layers with the same number of trainable parameters as the VGR Layers , and Original Model, which uses the original student model without distillation. During training, all distillation approaches use compact CoT data except Native SeqKD, which uses Native CoT responses.

**Additional Comparisons.** In addition to the six representative datasets reported in the main paper, we report the grounding performance (IoU@0.3 and IoU@0.5) on the remaining six datasets in Table 7.

Table 7: Grounding performance (IoU@0.3 and IoU@0.5) on the other six datasets for the three student models.

| Model | Approach | GQA | | VSR | | DocVQA | | InfoVQA | | DUDE | | SROIE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $IoU_{0.3}$ | $IoU_{0.5}$ | $IoU_{0.3}$ | $IoU_{0.5}$ | $IoU_{0.3}$ | $IoU_{0.5}$ | $IoU_{0.3}$ | $IoU_{0.5}$ | $IoU_{0.3}$ | $IoU_{0.5}$ | $IoU_{0.3}$ | $IoU_{0.5}$ |
| Qwen3-VL 2B | Token Swapping | 52.24 | 42.59 | 60.05 | 51.03 | 28.35 | 14.93 | 13.12 | 4.35 | 18.13 | 9.12 | 40.34 | 22.54 |
| | Few-Shot CoT | 52.80 | 43.65 | 61.73 | 53.09 | 28.49 | 15.40 | 12.53 | 4.42 | 18.13 | 9.56 | 41.67 | 22.35 |
| | Vanilla Layers | 57.31 | 47.27 | 66.37 | 56.31 | 37.13 | 20.44 | 20.80 | 7.49 | 26.07 | 14.71 | 43.56 | 25.76 |
| | VGG-CoT | 56.28 | 48.56 | 66.24 | 58.51 | 48.93 | 30.63 | 28.66 | 13.90 | 35.32 | 19.99 | 55.87 | 36.55 |
| Qwen2.5-VL 7B | Token Swapping | 53.10 | 41.60 | 62.50 | 49.23 | 11.84 | 4.01 | 7.79 | 2.06 | 9.56 | 3.17 | 23.48 | 7.77 |
| | Few-Shot CoT | 45.31 | 40.80 | 44.85 | 38.27 | 5.66 | 2.46 | 3.37 | 0.65 | 4.72 | 1.55 | 4.55 | 2.08 |
| | Vanilla Layers | 52.64 | 40.87 | 64.05 | 50.00 | 12.24 | 3.90 | 8.90 | 2.52 | 9.93 | 3.35 | 25.57 | 11.93 |
| | VGG-CoT | 56.58 | 46.54 | 65.85 | 55.67 | 17.68 | 6.58 | 15.44 | 4.68 | 14.65 | 5.59 | 32.58 | 18.56 |
| InternVL3 1B | Token Swapping | 43.19 | 31.65 | 57.09 | 41.62 | 10.18 | 2.68 | 1.31 | 0.26 | 2.61 | 0.50 | 1.52 | 0.19 |
| | Few-Shot CoT | 13.99 | 6.03 | 17.53 | 10.57 | 2.50 | 0.62 | 0.92 | 0.26 | 0.99 | 0.37 | 0.38 | 0.00 |
| | Vanilla Layers | 50.02 | 38.81 | 59.54 | 49.48 | 17.54 | 6.76 | 3.50 | 0.65 | 5.34 | 1.24 | 13.26 | 4.55 |
| | VGG-CoT | 52.54 | 43.55 | 63.02 | 54.25 | 24.52 | 11.10 | 6.54 | 1.70 | 8.94 | 2.98 | 11.55 | 2.84 |

**Statistical Significance Tests.** To assess whether the improvements of VGG-CoT over the baselines are statistically reliable rather than due to chance, we conduct paired two-sided Wilcoxon signed-rank tests on per-example scores matched by fingerprint, omitting unavailable scores and zero differences, and apply a Bonferroni-adjusted threshold of $\alpha = 0.005$ for the ten Qwen3-VL-2B metric comparisons (five comparison methods times the ANLS and token-level F1 metrics). As summarized in Table 8, VGG-CoT significantly outperforms every comparison approach on both metrics (all $p < 0.001$): versus Vanilla Layers, the mean differences are $+1.19$ (ANLS) and $+1.25$ (token F1); versus Token Swapping, they are $+6.96$ and $+6.73$; versus Native SeqKD, they are $+55.19$ and $+47.50$; versus Few-Shot CoT, they are $+16.94$ and $+14.56$; and versus Original Model, they are $+2.26$ and $+2.67$. All p-values are orders of magnitude below the Bonferroni-adjusted threshold of $0.005$, and even the smallest margin over the closest baseline (Vanilla Layers, about one percentage point) is highly significant. These results demonstrate that the advantages of VGG-CoT over all baselines are systematic and statistically reliable rather than artifacts of evaluation noise, supporting the effectiveness of our approach. The significance tests for the other student models, which follow the same protocol from one fixed-seed run, remain descriptive.

**Token Efficiency and Generation Length Distribution.** To compare the effects of native free-form CoT and our compact CoT on student model performance, we conduct experiments using Qwen3-VL-2B as the student model. Specifically, we train two instances of Qwen3-VL-2B, both with LoRA applied to all decoder layers, under identical settings except for the CoT data: one uses native CoT data and the other uses compact CoT data. After training, we evaluate both models on the CM-CoT validation set. Table 9 presents the per-category token-length statistics for both models. Compared with native CoT training, compact CoT training consistently reduces both the mean and variability of response lengths across all five task categories, yielding more concise and stable outputs.

## D Code and Data Availability

To facilitate reproducibility, the supplementary package is distributed alongside this PDF as a single ZIP archive containing two directories: `code`, which contains de-identified reference scripts

Table 8: Paired Wilcoxon signed-rank tests for Qwen3-VL-2B. Mean differences are VGG-CoT minus the comparison approach, p-values are two-sided, from paired per-example Wilcoxon tests, and evaluated with the Bonferroni-adjusted threshold $\alpha = 0.005$.

| Metric | VGG-CoT vs. | Mean $\Delta$ (pp) | $p$ |
|---|---|---|---|
| ANLS | Vanilla Layers | +1.19 | $< 0.001$ |
| | Token Swapping | +6.96 | $< 0.001$ |
| | Native SeqKD | +55.19 | $< 0.001$ |
| | Few-Shot CoT | +16.94 | $< 0.001$ |
| | Original Model | +2.26 | $< 0.001$ |
| Token F1 | Vanilla Layers | +1.25 | $< 0.001$ |
| | Token Swapping | +6.73 | $< 0.001$ |
| | Native SeqKD | +47.50 | $< 0.001$ |
| | Few-Shot CoT | +14.56 | $< 0.001$ |
| | Original Model | +2.67 | $< 0.001$ |

for training, evaluation, checkpoint merging, metric computation, and layer-wise analysis, and `data-sample`, which contains 2,000 example CM-CoT annotation records. The JSONL records preserve original relative image paths of the source images.

**Code organization.** The code directory is organized by the reproducibility checklist: preprocessing, training, evaluation, and analysis. The training and evaluation scripts used for the Qwen3-VL-2B VGR analysis are included in the corresponding directories. The reported layer curves are generated from the archived per-layer evaluation reports using the metric definitions described in Sec. B.

**Training and evaluation scripts.** The external package contains the formal Stage 1 grounding LoRA launcher for the selected contiguous VGR Layers, the Stage 2 all-layer reasoning launcher, the checkpoint-merging scripts, and the ANLS and token-level F1 computation script, which reads approach predictions from the de-identified `PREDICTIONS_ROOT` layout described in the package README. The source files preserve the recorded Qwen3-VL-2B training and evaluation control flow while replacing absolute paths, model locations, device identifiers, and other machine-specific values with placeholders or environment variables.

**Sample data and licensing.** The 2,000 CM-CoT annotation records are released for unrestricted academic and research use. The complete CM-CoT annotation file contains 120,000 records and is approximately 144 MB; due to its size, it will be provided after paper acceptance. The archive does not redistribute source images: the JSONL records retain relative paths to images that users must obtain from the original dataset projects. Those images remain subject to the individual source-dataset licenses, attribution requirements, and terms of use. The README in the supplementary package records these release conditions.

Table 9: Token-length statistics computed with the Qwen3-VL-2B tokenizer. Mean and standard deviation are reported in tokens.

| Task Category | Native | | Compact CoT | |
|---|---|---|---|---|
| | Mean | Std. | Mean | Std. |
| Visual Grounding | 346.43 | 1096.04 | 110.11 | 7.30 |
| General VQA | 378.70 | 1295.72 | 106.42 | 3.77 |
| Document Understanding | 445.18 | 1431.26 | 119.42 | 9.44 |
| Scene Text Understanding | 253.00 | 988.52 | 112.57 | 5.66 |
| Fine-grained Recognition | 274.53 | 174.09 | 112.31 | 5.27 |
| Overall | 349.35 | 1170.96 | 111.99 | 8.08 |

# E CASE STUDIES

In this section, we present twelve qualitative case studies illustrating how CM-CoT grounds its annotations in relevant visual evidence and avoids potential distractors across diverse VQA tasks. The questions are preserved verbatim from the source datasets, including any grammatical or OCR-like artifacts. Through these case studies, we further demonstrate that CM-CoT connects relevant visual evidence to the reasoning process, yielding visually grounded answers.

**Case 1: General VQA (V7W)**

**Question:** What number is above the doorway?

<find> Target: number. Region: above doorway. Why: displays 206. </find>

<reasoning> Task type: text reading. Answer format: exact text span. Visual anchor: crop shows number above doorway. Self-correction: not other text elements. </reasoning>

<result> **206** </result>



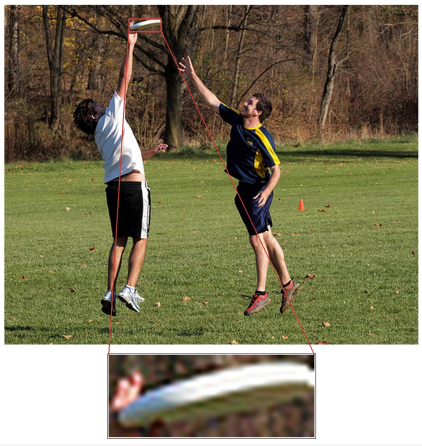

**Case 2: General VQA (GQA)**

**Question:** What is the man to the left of the player trying to catch?

<find> Target: frisbee. Region: flying object. Why: object the left man is catching. </find>

<reasoning> Task type: object identification. Answer format: object name. Visual anchor: crop shows the flying object the left man is catching. Self-correction: not a ball. </reasoning>

<result> **frisbee** </result>



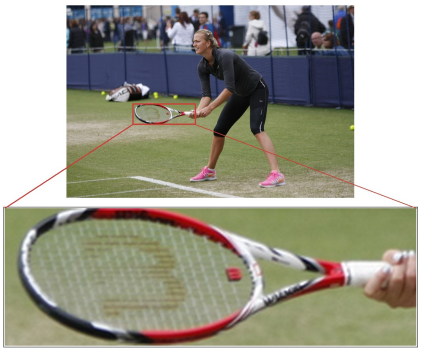
**Case 3: Visual Grounding (OpenImages)**

**Question:** What the standing woman holds?

<find> Target: tennis racket. Region: racket. Why: she holds it. </find>

<reasoning> Task type: object identification. Answer format: object name. Visual anchor: crop shows the racket she holds. Self-correction: not the bag/balls. </reasoning>

<result> **tennis racket** </result>



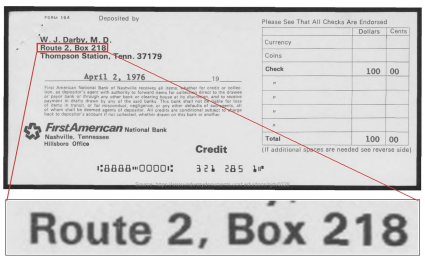
**Case 4: Document Understanding (DocVQA)**

**Question:** What is the BOX Number ?

<find> Target: BOX Number. Region: address line. Why: shows 'Box 218'. </find>

<reasoning> Task type: document field extraction. Answer format: number. Visual anchor: crop shows "Box 218" in address. Self-correction: not the route number (2). </reasoning>

<result> **218** </result>



**Case 5: Scene Text Understanding (TextCaps)**

**Question:** Which word on the sign suggests a connection to piracy?

<find> Target: SMUGGLERS. Region: bottom text. Why: word implies smuggling/piracy link. </find>

<reasoning> Task type: text reading. Answer format: exact text span. Visual anchor: crop shows "SMUGGLERS" at sign bottom. Self-correction: not "Marston's" (top text). </reasoning>

<result> **SMUGGLERS** </result>




**Case 6: Scene Text Understanding (TextCaps)**

**Question:** Which word in the sign implies a prohibition or restriction?

<find> Target: prohibition word. Region: "DON'T" text. Why: "DON'T" means do not, indicating restriction. </find>

<reasoning> Task type: text reading. Answer format: exact text span. Visual anchor: crop shows "DON'T", meaning do not. Self-correction: not "LET" or "BUY". </reasoning>

<result> **DON'T** </result>




**Case 7: Scene Text Understanding (TextVQA)**

**Question:** what is the player number on the left?

<find> Target: player number. Region: left player's jersey. Why: displays 23. </find>

<reasoning> Task type: text reading. Answer format: number. Visual anchor: crop shows left player's jersey number 23. Self-correction: not the right player's 5. </reasoning>

<result> **23** </result>




**Case 8: Scene Text Understanding (TextVQA)**

**Question:** what kind of stop is this?

<find> Target: stop type. Region: 4-WAY sign. Why: specifies 4-way stop. </find>

<reasoning> Task type: text reading. Answer format: exact text span. Visual anchor: crop shows 4-WAY sign, specifying stop type. Self-correction: not the main STOP sign. </reasoning>

<result> **4-way** </result>



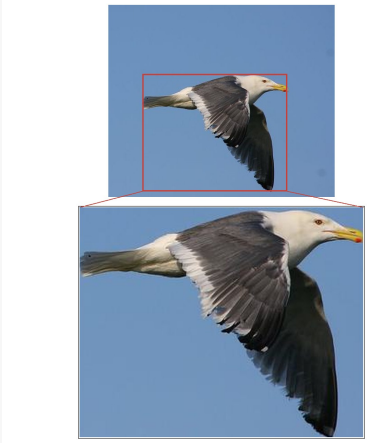
**Case 9: Fine-grained Recognition (CUB)**

**Question:** Does the bird in the picture have white crown and grey wing?

<find> Target: bird. Region: head and wing. Why: white crown and grey wing visible. </find>

<reasoning> Task type: object attribute. Answer format: yes/no. Visual anchor: crop shows white crown and grey wing. Self-correction: not tail or beak. </reasoning>

<result> **Yes** </result>



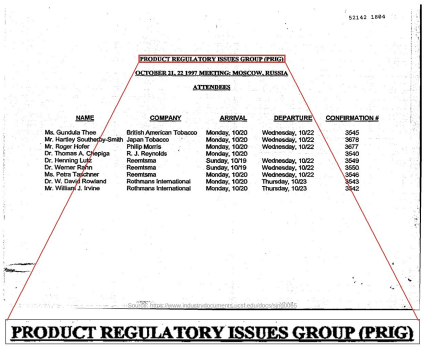
**Case 10: Document Understanding (DocVQA)**

**Question:** What is the fullform of PRIG?

<find> Target: PRIG full form. Region: title line. Why: defines PRIG as Product Regulatory Issues Group. </find>

<reasoning> Task type: text reading. Answer format: exact text span. Visual anchor: crop shows PRIG's full form in title. Self-correction: not the meeting details. </reasoning>

<result> **Product Regulatory Issues Group** </result>




**Case 11: Spatial Reasoning (VSR)**

**Question:** Does the oven contain the pizza?

<find> Target: pizza location. Region: pizza on tray. Why: pizza outside oven, answer No. </find>

<reasoning> Task type: spatial relation. Answer format: yes/no. Visual anchor: crop shows pizza on tray outside oven. Self-correction: oven opening is empty. </reasoning>

<result> **No** </result>



**Case 12: Infographic Understanding (InfographicsVQA)**

**Question:** Which state in America contributes to the least percent of Afro-American population?

<find> Target: state with least pop share. Region: Michigan's 14% blue bar. Why: 14% is smallest state value. </find>

<reasoning> Task type: document field extraction. Answer format: state name. Visual anchor: crop shows Michigan's 14% pop share, smallest state value. Self-correction: not Chicago (city, not state). </reasoning>

<result> **Michigan** </result>

