GroundCount: Grounding Vision-Language Models with Object Detection for Mitigating Counting Hallucinations

Boyuan Chen[1,2], Minghao Shao[1,2], Siddharth Garg[1], Ramesh Karri[1], Muhammad Shafique[2],
[1]Tandon School of Engineering, New York University, NY, USA
[2]eBRAIN Lab, Division of Engineering, New York University Abu Dhabi, UAE
{boyuan.chen, minghao.shao, sg175, rkarri, muhammad.shafique}@nyu.edu

***Abstract*—Vision Language Models (VLMs) exhibit persistent hallucinations in counting tasks, achieving substantially lower accuracy than other visual reasoning tasks. This phenomenon persists even in state-of-the-art reasoning-capable VLMs. CNN-based object detection models (ODMs) such as YOLO, by contrast, excel at spatial localization and instance enumeration with negligible computational overhead. We propose *GroundCount*, a framework that augments VLMs with explicit spatial grounding from ODMs to mitigate counting hallucinations. Our prompt-based augmentation achieves up to 80.6% counting accuracy on Ovis2.5-2B (a 5.9pp improvement), with gains reaching 7.7pp on Molmo2-4B across evaluated architectures, while reducing inference time by up to 24% for models prone to hallucination-driven reasoning loops such as Molmo2-4B, whose inference time drops from 8.0s to 6.1s. Comprehensive ablation studies reveal that positional encoding is a critical component, with its effect being architecture-specific: beneficial for most architectures yet detrimental for others regardless of model scale. Confidence scores, by contrast, offer marginal value across most architectures, improving performance in three of five evaluated models with modest gains of 0.2–1.1pp. We further evaluate feature-level fusion architectures and find that explicit symbolic grounding via structured prompts consistently outperforms implicit feature fusion, despite the latter employing sophisticated cross-attention mechanisms. Overall, our approach yields consistent improvements across four of five evaluated VLM architectures, with per-model gains ranging from 5.9 to 7.7pp; the remaining architecture exhibits comparatively limited gains, likely attributable to its iterative reflection mechanism partially overriding structured prompt inputs. We open-source our code at https://github.com/BoyuanChen99/GroundCount.**

I. Introduction

Vision-Language Models (VLMs) have demonstrated remarkable progress across a wide range of multimodal tasks [1], benefiting from large-scale training and fine-tuning on massive image-text corpora. This joint learning paradigm significantly enhances both contextual reasoning and instruction-following abilities [2]. Despite such advances, VLMs exhibit *hallucinations* by generating object attributes, entities, or spatial relations inconsistent with the visual input [3], [4].

A growing body of work attributes this issue to an imbalanced cross-modal attention mechanism: the language decoder often over-attends to textual priors while under-utilizing visual tokens [5]–[7]. Consequently, several methods have been proposed to mitigate hallucination, either through decoding-level adjustments [8]–[10] or training-level regularization [11], [12]. Motivated by layer-level analysis showing that hallucination primarily occurs when later layers deviate from correct logits, another direction proposes steering strategies [13]–[15] that preserve the correct momentum from earlier layers. These approaches have yielded measurable improvements on popular VLMs, including LLaVA-v1.5 [16], LLaVA-v1.6 [2], InstructBLIP [17], MiniGPT-2 [18], and mPLUG-Owl2 [19].

Recent work jointly trains vision encoders and language decoders on diverse multimodal datasets [20], and incorporates reinforcement learning to enhance reasoning capabilities [21]–[23]. These efforts have led to significantly stronger multimodal representations and improved factual alignment. However, even state-of-the-art (sota) VLMs continue to exhibit systematic hallucination in *counting* tasks. As we demonstrate in Section III and Table I, counting consistently remains the lowest-accuracy task across all evaluated VLMs (excluding sentiment, which is relatively subjective), with accuracies ranging from 64.1% to 74.7%. This is substantially lower than other visual reasoning tasks such as object recognition (70.0% to 89.6%) and attribute identification (79.9% to 87.0%). Recent research also supports this finding [24], [25].

The persistence of counting hallucinations in reasoning-capable VLMs presents a unique challenge. Unlike earlier VLMs where attention steering and decoding adjustments proved effective, modern reasoning VLMs already incorporate reflection mechanisms [20], [22], [23] that direct attention to visual modalities. Furthermore, the iterative reasoning process makes it difficult to apply layer-level vector steering, as there is no single "correct" token to steer toward during multi-step reasoning. This suggests that existing hallucination mitigation strategies, while effective for simpler VLMs, are insufficient for addressing systematic failures in compositional tasks like counting.

On the other hand, counting the number of occurrences of an object is nearly trivial for object detection models (ODMs), such as YOLO [26] and DETR [27]. These models achieve very high accuracy in object detection while requiring minimal inference time compared to auto-regressive VLMs. Moreover, ODMs provide structured outputs including bounding boxes, confidence scores, and class labels. This information explicitly

addresses the spatial and compositional reasoning needed for accurate counting. This observation suggests a complementary approach: rather than attempting to fix VLM attention mechanisms or reasoning processes, we can augment VLMs with explicit grounding information from specialized object detection models.

Our contributions are as follows:

- We provide a comprehensive analysis of sota VLMs in counting tasks, demonstrating that counting remains systematically the lowest-accuracy task across all evaluated models despite advances in reasoning capabilities.
- We introduce **GroundCount**, an object-detection-driven augmentation pipeline that increases VLM counting accuracy by 5.9 to 7.7pp across four of five evaluated models (up to 80.6% on the PhD benchmark) with negligible memory overhead and reduced inference time for stronger models.
- We conduct three groups of ablation studies analyzing the impact of different ODM outputs (confidence scores, positional encoding, detection thresholds) on VLM performance, providing insights into which ODM output components are responsible for accuracy gains.
- We evaluate a fusion architecture that combines VLMs with ODMs at the feature level through fine-tuning, analyzing the trade-offs between prompt-based augmentation and architectural integration.

II. Related Works

VLM Hallucination in Counting Tasks: Vision-Language Models exhibit persistent hallucinations in counting tasks despite advances in reasoning capabilities. One major hypothesis attributes this to vision transformers (ViTs) having difficulties noticing fine-grained image details [28], [29], as ViTs process images through global patch attention rather than the local receptive fields that facilitate spatial instance discrimination [28], [30]. Recent studies [24], [31] systematically demonstrate that counting remains a fundamental challenge for VLMs, with accuracy substantially lower than other visual reasoning tasks.

VLM Hallucination Mitigation Methods: Existing approaches to mitigate VLM hallucinations operate at multiple levels. Decoding-level methods adjust the generation process through contrastive decoding [12], layer contrasting [7], or attention-guided decoding [9]. Training-level approaches incorporate regularization [11] or specialized fine-tuning objectives. More recent steering strategies [14], [15] preserve correct activations from earlier layers to prevent hallucination drift. However, these methods primarily target general hallucinations and show limited effectiveness on systematic compositional failures like counting, particularly in reasoning-capable VLMs that already employ reflection mechanisms.

CNN–Transformer Fusion and Cross-Modal Grounding: DETR [27] pioneered combining CNN backbones with transformer decoders, demonstrating how structured CNN features can enhance transformer reasoning; subsequent work explores broader fusion strategies [32], [33]. Multimodal fusion requires careful cross-modal alignment, for which cross-attention [34], [35] and Feature-wise Linear Modulation (FiLM) [36] have proven effective building blocks. Our fusion architecture (Plans B and C) extends these ideas to grounding VLM patch tokens with CNN-based object detection features, bridging the representational gap between local CNN outputs and global ViT patch embeddings.

Object Detection for Visual Reasoning: Object detection models provide structured, localized visual information that can ground higher-level reasoning. YOLO [37] and its variants achieve real-time detection with high accuracy through efficient CNN architectures, excelling at spatial localization and instance counting — precisely where VLMs struggle. Prior work has explored using detected objects as input to VQA systems [38], [39], but typically through learned feature fusion during training rather than explicit textual grounding [40], [41]. Our prompt-based approach (Plan A) uniquely leverages ODM outputs as interpretable, structured prompts that augment VLMs without architectural modification, while Plans B and C extend this with feature-level integration for deeper cross-modal grounding.

III. Analysis - Hallucination in Counting Tasks

We first validate that state-of-the-art VLMs systematically exhibit counting hallucinations through controlled experiments. We employ consistent benchmark and model selections across all evaluations in this work.

A. VQA Benchmark

We employ the PhD benchmark [42] to evaluate vision-language model capabilities. The benchmark comprises 33,688 visual question-answer (VQA) pairs across 16,844 images with binary (yes/no) ground-truth labels. Questions are categorized

TABLE I: Correctness rates (%) of different VLMs on the PhD benchmark using greedy decoding, with explicit thinking mode enabled when available (i.e., the output begins with a prior reasoning segment enclosed by special tokens such as "<think>...</think>"). The PhD benchmark consists of questions from five task types listed in the **Task** column on the left. For each subset (excluding sentiment), the task with the lowest correctness rate is highlighted with a light-red background. Results across all four evaluation settings are shown; baseline corresponds to standard VQA without adversarial context.

VLM	**Molmo2-4B**				**Ovis2.5-2B**				**R-4B**				**Qwen3-VL-2B**				**InternVL3.5-1B**			
Task	base	sec	icc	ccs	base	sec	icc	ccs	base	sec	icc	ccs	base	sec	icc	ccs	base	sec	icc	ccs
object	89.6	79.9	67.1	89.2	88.5	85.1	84.0	84.6	88.0	84.5	77.0	90.4	87.8	84.0	81.5	83.1	70.0	66.7	64.8	82.0
attribute	85.1	43.0	15.5	93.9	86.2	76.8	75.3	89.2	87.0	71.4	58.3	92.7	86.5	74.4	66.5	88.1	79.9	68.9	64.6	80.9
positional	80.4	66.9	47.1	89.1	80.9	73.0	65.3	88.6	80.4	71.0	55.2	94.1	76.9	69.7	63.6	86.8	70.6	63.9	55.1	85.9
counting	67.2	63.1	34.2	87.9	74.7	67.6	52.0	79.8	73.9	69.0	42.7	87.9	70.6	66.4	56.3	83.1	64.1	57.9	50.0	85.5
sentiment	67.5	52.8	32.9	93.6	68.4	59.6	59.2	94.9	67.3	54.0	45.6	96.2	68.1	47.2	44.1	91.0	62.9	51.6	51.7	83.3

into five task types: *attribute*, *counting*, *object*, *positional*, and *sentiment*. To assess robustness against misleading information, the benchmark provides three challenge variants: (1) *sec* (specious context) with plausible but potentially misleading descriptions, (2) *icc* (incorrect context) where VLMs receive explicitly incorrect textual descriptions, and (3) *ccs* (contradictory common sense) featuring 1,506 VQA pairs on 753 AI-generated images that violate common-sense expectations. Compared to earlier benchmarks such as POPE [43] and CHAIR [44], PhD provides larger sample sizes and more granular task categorization.

B. Selection of VLM and ODM

We evaluate five state-of-the-art open-source VLMs with parameters not exceeding 4B: R-4B [21], Ovis2.5-2B [23], Qwen3-VL-2B-Thinking (abbreviated as Qwen3-VL-2B for the rest of the paper) [45], InternVL3.5-1B [22], and Molmo2-4B [46]. This selection spans diverse architectural designs, with several incorporating advanced reasoning mechanisms such as reinforcement learning (R-4B) and iterative reflection (InternVL3.5). All models are loaded in `float32` precision. We allocate generous computational budgets of 1,024 tokens for both output generation and thinking steps, using greedy decoding to ensure reproducibility. For object detection, we utilize YOLOv13x [26] as our primary ODM, selected for its state-of-the-art accuracy and minimal inference latency.

C. Results and Discussion

Table I presents the comprehensive evaluation results. Our findings clearly demonstrate: **Counting consistently remains the lowest-accuracy task in standard evaluation. Notably, attribute identification exhibits steeper degradation under adversarial contexts in higher-capacity models, suggesting differential fragility across task types.**

In the base evaluation setting, counting accuracy ranges from 64.1% to 74.7% (mean: 70.1%), representing a substantial gap compared to other visual reasoning tasks: object recognition achieves 70.0%–89.6% (mean: 84.8%), attribute identification reaches 79.9%–87.0% (mean: 84.9%), and positional reasoning attains 70.6%–80.9% (mean: 77.8%). This represents an average accuracy deficit of approximately 14.7–14.8pp compared to object recognition and attribute identification, respectively.

Performance degradation under misleading contexts further reveals the fragility of counting abilities. Under *sec* and *icc* conditions, counting accuracy drops by 4.1–7.1pp and 14.1–33.0pp, respectively, among the most consistent degradations across all five architectures and task categories. This suggests that VLM counting mechanisms are particularly susceptible to distraction from textual priors, consistent with hypotheses about imbalanced cross-modal attention [5].

Model-Specific Patterns. Counting failures are not simply correlated with parameter count or training paradigm: Molmo2-4B achieves only 67.2% baseline accuracy despite its 4B parameter scale, while the smaller Ovis2.5-2B reaches

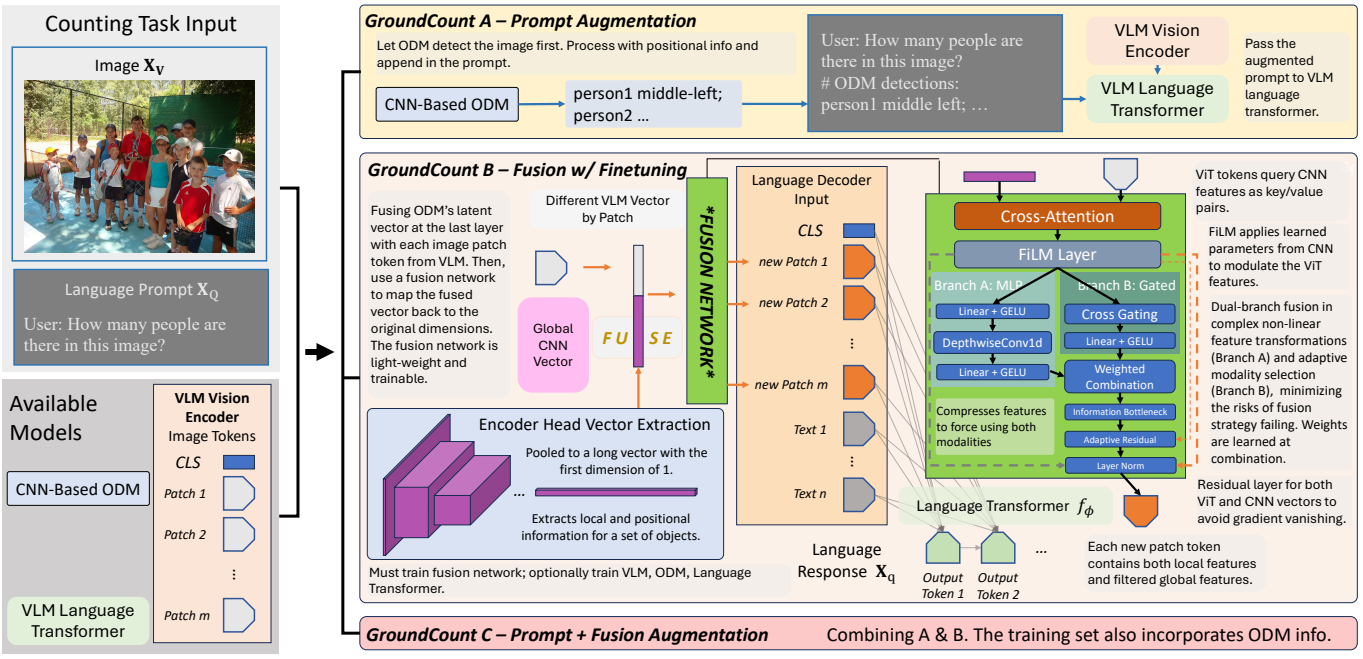

Fig. 1: Structural overview of three strategies in our proposed fusion framework - A, B and C. In **GroundCount A**, we run inference with ODM on the image, and then include its output in the VLM prompt. In **GroundCount B**, we fuse the VLM and ODM on the visual patch latent vector using a light-weight network. To ensure correct information delivery, we finetune the network with our original counting task mutation from COCO. The fusion block is required to be trained; Other modules - VLM, ODM, and language transformer - are optionally frozen. **GroundCount C** incorporates both plans by including both prompt-level information and architectural-level integration. The training data also includes ODM detections in the textual input.

74.7%. Under *sec*, counting accuracy degrades most sharply in Ovis2.5-2B (7.1pp), followed by InternVL3.5-1B (6.2pp), whose disproportionate sensitivity relative to its model size suggests its iterative reflection mechanism requires greater capacity to filter spurious textual priors.

These findings suggest the problem lies not in reasoning depth but in fundamental spatial-semantic integration. Rather than refining attention mechanisms within existing VLM paradigms, we augment VLMs with explicit grounding from specialized object detection models that excel precisely where VLMs fail.

IV. Methodology - Augmenting VLMs with Object Detection Model

We propose GroundCount, a framework that augments VLMs with explicit spatial grounding from object detection models to mitigate counting hallucinations. Our approach operates on the observation that CNN-based ODMs excel at spatial localization and instance counting, precisely where VLMs exhibit systematic failures. We present three implementation strategies with different computational trade-offs.

A. GroundCount A: Prompt-Based ODM Augmentation

The most straightforward approach augments VLM prompts with structured textual descriptions of ODM detections. As illustrated in Figure 2, we first pass the input image through YOLOv13x to obtain bounding boxes, class labels, and confidence scores, which are then converted into natural language prompts.

Spatial Encoding: We discretize image space into a 3×3 grid (upper/middle/lower × left/center/right) and assign each detection to a grid cell based on its bounding box center coordinates. This coarse spatial encoding preserves relative positioning information while remaining interpretable for language models.

Object Sequencing: Detections are ordered by (1) horizontal position (left-to-right), then (2) vertical position (lower-to-upper). This consistent ordering enables VLMs to maintain spatial coherence when processing multiple instances of the same object class.

Prompt Construction: For each detection, we generate a string in the format: "`[class] [index] [position]: [confidence]`", where the index distinguishes multiple instances of the same class. The complete ODM output is appended to the original user prompt as structured context (see Figure 2).

This approach requires no architectural modifications or task-specific training, making it plug-and-play across diverse VLM architectures. The computational overhead is minimal: YOLOv13x inference (∼0.1s) is negligible compared to VLM auto-regressive decoding (8–70s depending on model architecture and reasoning complexity), and sometimes reduces total inference time by preventing hallucination-induced reasoning loops, particularly for stronger models (Figure 4).

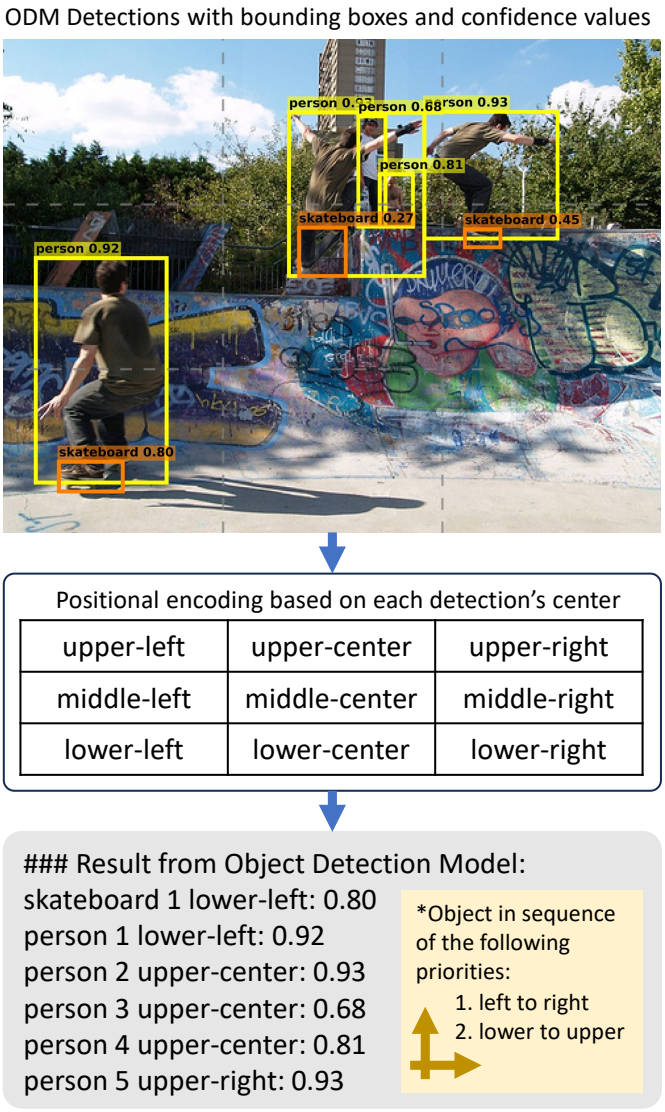

Fig. 2: Our pipeline of converting ODM outputs to descriptive text. The image is #000000000077.jpg from COCO-train2017, showing 5 young people skateboarding. The bounding boxes (bbox) come from YOLOv13x's detection: yellow ones are `person` objects; orange ones are `skateboard` objects. The location of each object is determined by the center of their corresponding bbox. Two skateboard objects were not included due to low confidence.

B. GroundCount B: Feature-Level Fusion Architecture

While prompt augmentation proves effective, it relies on the VLM's ability to correctly interpret textual descriptions of spatial information. To enable direct feature-level grounding, we propose a fusion architecture that integrates CNN features from the ODM with ViT patch tokens from the VLM's vision encoder.

Architecture Overview: As shown in Figure 1, our fusion network operates between the VLM's vision encoder and language decoder. For each ViT patch embedding $\mathbf{p}_i \in \mathbb{R}^{d_{\text{vit}}}$, we extract the corresponding spatial region from the ODM's final convolutional layer, yielding a local CNN feature $\mathbf{c}_i \in \mathbb{R}^{d_{\text{cnn}}}$.

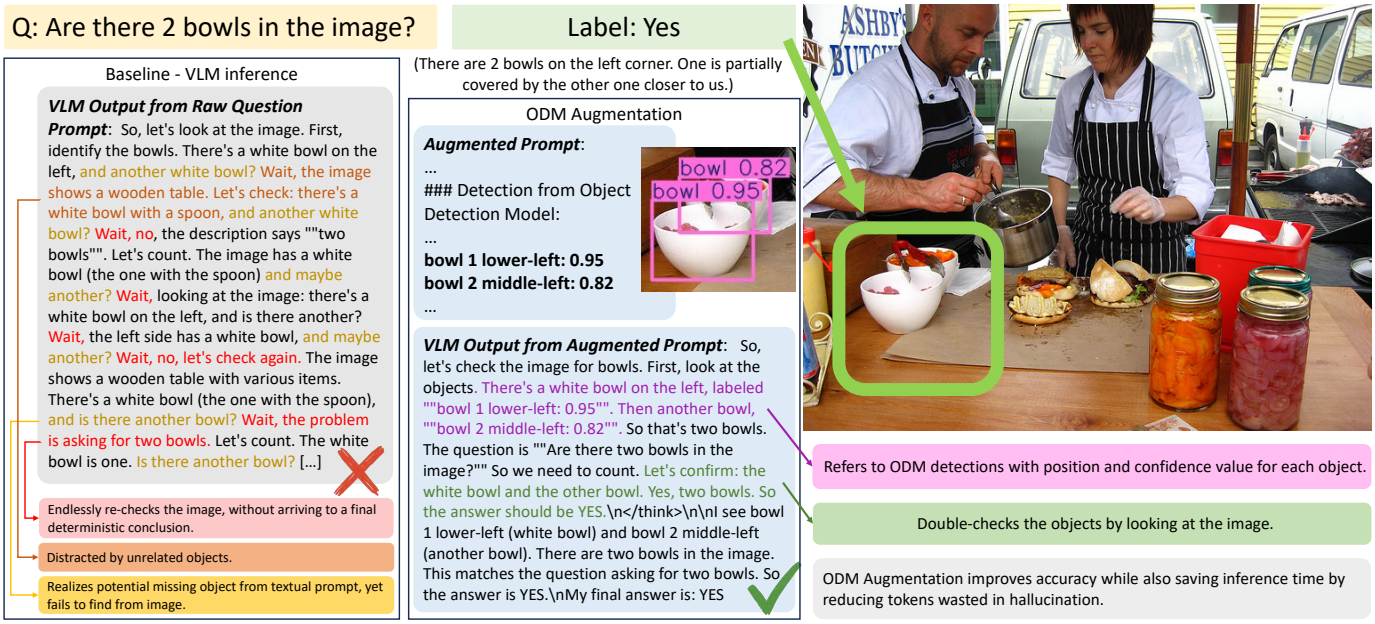

Fig. 3: Illustration with real evaluation example for ODM prompt augmentation in GroundCount A. The image is #000000189241.jpg from COCO-val2014, which is used for a counting question in our selected benchmark, PhD. The tested VLM is Qwen3-VL-2B-Thinking. The question asks for a correctness judgment on the number of bowls in the image. There are indeed 2 bowls in the image, with one bowl partially covered by another. The baseline VLM, whose output is showcased in the left frame, fails to find the second bowl and re-thinks iteratively, exhibiting behaviors of hallucination. On the other hand, with information from the object detection model (ODM) appended in the prompt, the VLM successfully finds the second bowl with a double-check.

We additionally extract a global CNN feature vector $\mathbf{g} \in \mathbb{R}^{d_{\text{cnn}}}$ via adaptive pooling.

Dual-Branch Fusion: The fusion network employs two parallel branches to integrate multimodal features:

Branch A (Feature Transformation): Applies Feature-wise Linear Modulation (FiLM) [36] to enable CNN features to adaptively modulate ViT representations:

$$\mathbf{h}_i^A = \text{FiLM}(\mathbf{p}_i, \mathbf{c}_i) = \boldsymbol{\gamma}_i \odot \mathbf{p}_i + \boldsymbol{\beta}_i$$

where $\boldsymbol{\gamma}_i, \boldsymbol{\beta}_i = \text{MLP}(\mathbf{c}_i)$ are learned affine parameters.

Branch B (Selective Attention): Uses cross-attention for ViT patches to attend over relevant CNN features:

$$\mathbf{h}_i^B = \text{CrossAttn}(\mathbf{p}_i, [\mathbf{c}_i, \mathbf{g}])$$

The branches are combined via learned gating: $\mathbf{h}_i = \alpha_i \mathbf{h}_i^A + (1 - \alpha_i)\mathbf{h}_i^B$, where $\alpha_i = \sigma(\text{MLP}([\mathbf{p}_i, \mathbf{c}_i]))$. This dual-branch design provides robustness. Branch A enforces strong CNN influence through multiplicative modulation, while Branch B enables adaptive selection. This minimizes risk if one fusion strategy proves suboptimal.

Information Bottleneck: Following the dual-branch fusion, we apply dimensionality reduction to enforce multimodal information integration. This bottleneck prevents the model from trivially bypassing fusion by relying solely on one modality.

Training Data Preparation: We construct training data from the COCO train2017 dataset [47] using ground-truth object annotations. For each image, we apply the same spatial encoding and sequencing rules used for ODM detections in Plan A. Specifically, each ground-truth bounding box is assigned to a 3×3 grid cell based on its center coordinates, and objects are ordered left-to-right, then lower-to-upper.

The model is trained to generate structured spatial descriptions in the format: "`[class] [index] in [position]`" for each object instance. For example, given an image with multiple birds, the target output would be: *"bird 1 in upper-left; bird 2 in middle-center; bird 3 in lower-right"*. This format mirrors the ODM detection output structure, enabling the fusion model to learn consistent spatial grounding patterns that align with both ground-truth annotations and runtime ODM predictions.

Training Strategies: We evaluate four training configurations (Table II):

- **B.1**: Train fusion network only (frozen VLM, frozen ODM)
- **B.2**: Train fusion network + fine-tune VLM and ODM
- **B.3**: Train fusion network + fine-tune language decoder only
- **B.4**: Train fusion network + fine-tune VLM, ODM, and language decoder

C. GroundCount C: Combined Prompt and Fusion

Our final approach combines both strategies: structural fusion (Plan B.4) with prompt augmentation (Plan A). This hybrid design provides complementary benefits. Feature-level fusion enables implicit spatial grounding, while textual prompts offer explicit, interpretable object counts. The combined approach aims to leverage both implicit feature integration and explicit symbolic reasoning.

D. Implementation Details

All experiments use YOLOv13x as the object detection model with default confidence threshold 0.5. For Plan A, we filter detections below this threshold before prompt construction. VLMs are evaluated with greedy decoding, 1024-token output budget, and float32 precision. Fusion network training uses AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$), learning rate 2×10^{-5} with cosine annealing, batch size 1, and training with a maximum training budget of 40k steps; best checkpoints selected per Table II. All experiments are conducted on NVIDIA A100 GPUs with 80GB memory.

V. Results & Analysis

A. Main Results: Comparing GroundCount Strategies

Table II presents a comprehensive comparison of our proposed GroundCount strategies on the PhD benchmark's counting subset, evaluated on Ovis2.5-2B.

Prompt Augmentation (Plan A) achieves the highest counting accuracy at 80.6%, a 5.9pp improvement over the 74.7% baseline. The negligible computational overhead of YOLOv13x inference ($\sim$0.1s) introduces minimal latency for Ovis2.5-2B ($+$0.4s), while for models prone to hallucination-driven reasoning loops (e.g., Molmo2-4B and R-4B) it actively reduces total inference time; for instance, Molmo2-4B decreases from 8.0s to 6.1s (-24%). This reduction is attributable to VLMs generating fewer tokens when supplied with explicit grounding information, an effect particularly pronounced in complex chain-of-thought scenarios (Figure 3).

Fusion Architectures (Plan B) demonstrate varied outcomes. Training only the fusion network (B.1) yields minimal improvement (75.2%), while jointly fine-tuning VLM and ODM without the language decoder (B.2) degrades performance to 72.7%, suggesting that unconstrained joint fine-tuning disrupts the pretrained VLM's representations. Extending fine-tuning to include the language decoder (B.4) recovers this degradation, achieving 78.0% accuracy — a meaningful 3.3pp gain over baseline, though still below Plan A — while also reducing inference time to 7.7s compared to Plan A's 10.4s.

Combined Strategy (Plan C) achieves the lowest inference time among all schemes that exceed baseline accuracy (4.8s), with 78.2% accuracy. However, accuracy remains below Plan A alone, suggesting the fusion network introduces noise that partially counteracts explicit textual grounding.

These findings reveal a critical insight: for counting tasks, *explicit symbolic grounding via prompts outperforms implicit feature-level fusion*. This aligns with recent work emphasizing interpretable intermediate representations in multimodal reasoning [35], and likely reflects the fundamental representational gap between CNN spatial features and ViT global patch embeddings.

B. Ablation Study: Decomposing ODM Information

Figure 4 presents ablation results across five VLM families, systematically removing different components of ODM information.

Confidence Scores: Excluding confidence scores (NoConfidence) yields mixed results across architectures. While Ovis2.5-2B shows no accuracy change (0.0pp) and Molmo2-4B marginally improves (0.1pp), the remaining three models show slight decreases of 0.2 to 1.1pp, suggesting confidence values provide marginal and mixed benefit for most architectures.

Positional Encoding: Removing positional encoding (NoPosition) causes clear degradation in Molmo2-4B (-1.8pp) and InternVL3.5-1B (-1.4pp), while Qwen3-VL-2B *improves* by 0.6pp without positional information and R-4B sees a marginal decrease of 0.1pp; Ovis2.5-2B shows a negligible gain of 0.1pp. This pattern is architecture-specific rather than correlated with model scale, indicating inconsistent interpretation of spatial encodings across architectures.

Detection Threshold: Lowering the confidence threshold from 0.5 to 0.3 (LowThreshold) uniformly harms performance, with accuracy decreasing 5.7–16.9pp relative to Full ODM augmentation and inference time increasing 4.0–58.3s relative to the no-augmentation baseline. Degradation is most severe in Molmo2-4B (16.9pp) and Qwen3-VL-2B (15.6pp), consistent with stronger ODM reliance amplifying sensitivity to false-positive injections. This suggests that *detection precision is substantially more critical than recall* for effective counting augmentation.

Model-Specific Patterns: Molmo2-4B uniquely supports a pointing mode where bounding boxes are overlaid on the input image. This visual grounding (71.1%) underperforms textual ODM augmentation (74.9%), suggesting Molmo's architecture processes structured textual descriptions more effectively than visual annotations.

C. ODM-Only Baseline

Running YOLOv13x alone achieves 72.8% accuracy in 0.1s, an important reference point showing the ODM already captures most counting information. When properly grounded with ODM outputs, the strongest VLMs (Plan A: Ovis2.5-2B at 80.6%) exceed ODM-only performance by 7.8pp, demonstrating that VLMs contribute valuable contextual reasoning when supplied with explicit spatial priors.

D. Cross-Architecture Analysis: Consistent Gains and Divergent Responses

The benefits of ODM augmentation show significant variation across architectures. Four of five evaluated models exhibit substantial accuracy improvements with Plan A: Molmo2-4B (7.7pp), Ovis2.5-2B (5.9pp), R-4B (6.7pp), and Qwen3-VL-2B (6.5pp).

However, InternVL3.5-1B presents a notable exception, showing comparatively modest improvement (64.1% $\rightarrow$ 67.5%). Its iterative reflection mechanisms likely partially override structured prompts; the slight decreases under NoPosition (-1.4pp) and NoConfidence (-0.2pp) ablations suggest all ODM output components contribute to its gains nonetheless. This divergence confirms that while counting hallucinations stem from fundamental spatial-semantic integration

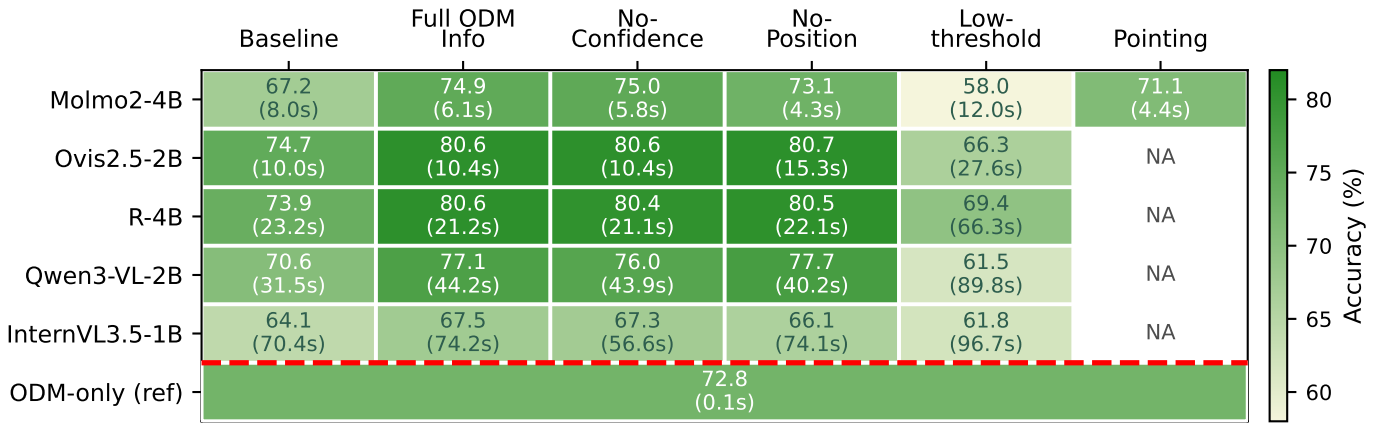

Fig. 4: Results of GroundCount A across all model families and including ablation studies. Each block contains the accuracy and average inference time for that group of experiment on the PhD counting subset. The bottom row marks the result of running the object detection model only. The right-most column records the special pointing mode for Molmo2 model only.

limitations, augmentation effectiveness depends critically on architectural compatibility.

TABLE II: Performance comparison of different GroundCount schemes - prompt augmentation (Plan A), architectural fusion with training (Plan B), and both (Plan C). Plan A achieves the highest accuracy (80.6%). Among schemes exceeding baseline accuracy, Plan C achieves the lowest inference time (4.8s), with Plan B.4 also improving over Plan A (7.7s vs. 10.4s); both are attributable to language transformer fine-tuning, with Plan C being additionally accelerated by joint prompt-fusion training reducing hallucination-driven reasoning loops.

Training Scheme	Acc(%)	Time(s)	Best Steps
Baseline	74.7	10.0	NA
Plan A - ODM Prompt Augmentation	80.6	10.4	NA
Plan B.1 - Fusion only	75.2	17.8	10k
Plan B.2 - Fusion + VLM, ODM	72.7	14.6	5k
Plan B.3 - Fusion + LangTrans	71.4	4.3	5k
Plan B.4 - Fusion + VLM, ODM + LangTrans	78.0	7.7	5k
Plan C - Plan A + Plan B.4	78.2	4.8	5k

VI. Conclusion

We present GroundCount, a framework that mitigates counting hallucinations in VLMs through explicit grounding from object detection models. Our evaluation across five state-of-the-art VLMs shows counting remains the lowest-accuracy task (64.1 to 74.7%, excluding sentiment). Our prompt-based augmentation (Plan A) achieves 80.6% counting accuracy on Ovis2.5-2B, a 5.9pp improvement over baseline, with negligible latency overhead; for models prone to hallucination-driven reasoning loops (e.g., Molmo2-4B), it additionally reduces inference time by up to 24%. Ablation studies reveal positional encoding benefits most architectures (up to 1.8pp for Molmo2-4B, 1.4pp for InternVL3.5-1B), yet proves detrimental for Qwen3-VL-2B (-0.6pp), suggesting architecture-specific sensitivity to spatial encodings. Critically, we find that explicit symbolic grounding via structured prompts outperforms feature-level fusion on Ovis2.5-2B, while prompt-based augmentation achieves consistent gains across four of five architectures. This indicates counting failures stem from fundamental spatial-semantic integration limitations rather than architecture-specific deficiencies. The gap between ODM-only performance (72.8%) and augmented VLM performance (80.6%) confirms that, when anchored by reliable spatial priors, VLMs contribute non-trivial contextual reasoning beyond what object detection alone provides. Future work includes extending fusion pre-training beyond 40k steps, exploring transformer-based ODMs (e.g., DETR, Grounding DINO) to reduce the CNN–ViT representational gap, and developing architecture-specific augmentation strategies for models with iterative reflection mechanisms.

Impact Statement

This work improves VLM reliability in counting tasks for accessibility and inventory applications. Improved counting could amplify surveillance risks; responsible deployment with proper consent is encouraged.

Acknowledgments

This work has been supported in parts by the NYUAD Center for Cyber Security (CCS), funded by Tamkeen under the NYUAD Research Institute Award G1104. Experiments are performed with NYUAD Jubail High Performance Computing (HPC).

References

[1] M. Shao *et al.*, "Survey of different large language model architectures: Trends, benchmarks, and challenges," *IEEE Access*, 2024.

[2] H. Liu *et al.*, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024.

[3] L. Zhao *et al.*, "Mitigating object hallucination in large vision-language models via image-grounded guidance," in *Forty-second International Conference on Machine Learning*, 2025.

[4] X. Zou *et al.*, "Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models," *The Forty-second International Conference on Machine Learning (ICML)*, 2025.

[5] S. Leng *et al.*, "Mitigating object hallucinations in large vision-language models through visual contrastive decoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 13 872–13 882.

[6] X. Wang *et al.*, "Mitigating hallucinations in large vision-language models with instruction contrastive decoding," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 15 840–15 853.

[7] Y.-S. Chuang *et al.*, "Dola: Decoding by contrasting layers improves factuality in large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[8] X. Xu *et al.*, "Mitigating hallucinations in multi-modal large language models via image token attention-guided decoding," 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)*. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 1571–1590.

[9] S. Yin *et al.*, "Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal llms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 16 520–16 530.

[10] Z. Jiang *et al.*, "Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2025.

[11] Y. Wu *et al.*, "Antidote: A unified framework for mitigating lvlm hallucinations in counterfactual presupposition and object perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 14 646–14 656.

[12] Z. Yang *et al.*, "Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 10 610–10 620.

[13] Z. Li *et al.*, "The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering," in *Forty-second International Conference on Machine Learning*, 2025.

[14] K. Wang *et al.*, "DAMO: Decoding by accumulating activations momentum for mitigating hallucinations in vision-language models," in *The Thirteenth International Conference on Learning Representations*, 2025.

[15] C. Wang *et al.*, "MLLM can see? dynamic correction decoding for hallucination mitigation," in *The Thirteenth International Conference on Learning Representations*, 2025.

[16] H. Liu *et al.*, "Visual instruction tuning," in *NeurIPS*, 2023.

[17] W. Dai *et al.*, "Instructblip: Towards general-purpose vision-language models with instruction tuning," 2023.

[18] J. Chen *et al.*, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," 2023.

[19] Q. Ye *et al.*, "mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024.

[20] J. Zhu *et al.*, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," 2025. [Online]. Available: https://arxiv.org/abs/2504.10479

[21] Q. Yang *et al.*, "R-4b: Incentivizing general-purpose auto-thinking capability in mllms via bi-mode annealing and reinforce learning," 2025. [Online]. Available: https://arxiv.org/abs/2508.21113

[22] W. Wang *et al.*, "Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency," 2025. [Online]. Available: https://arxiv.org/abs/2508.18265

[23] S. Lu *et al.*, "Ovis2.5 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2508.11737

[24] A. Vo *et al.*, "Vision language models are biased: Counting legs of an animal is surprisingly hard," in *2nd AI for Math Workshop @ ICML 2025*, 2025.

[25] X. Guo *et al.*, "Your vision-language model can't even count to 20: Exposing the failures of vlms in compositional counting," 2025.

[26] M. Lei *et al.*, "Yolov13: Real-time object detection with hypergraph-enhanced adaptive visual perception," 2025.

[27] N. Carion *et al.*, "End-to-end object detection with transformers," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 213–229.

[28] W. Park *et al.*, "SECOND: Mitigating perceptual hallucination in vision-language models via selective and contrastive decoding," in *Forty-second International Conference on Machine Learning*, 2025.

[29] M. Raghu *et al.*, "Do vision transformers see like convolutional neural networks?" in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021.

[30] M. Naseer *et al.*, "Intriguing properties of vision transformers," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021.

[31] S. Sengupta *et al.*, "Can vision-language models count? a synthetic benchmark and analysis of attention-based interventions," 2025.

[32] Z. Liu *et al.*, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11 976–11 986.

[33] Z. Dai *et al.*, "Coatnet: marrying convolution and attention for all data sizes," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021.

[34] J.-B. Alayrac *et al.*, "Flamingo: a visual language model for few-shot learning," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Curran Associates Inc., 2022.

[35] J. Li *et al.*, "Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[36] E. Perez *et al.*, "Film: visual reasoning with a general conditioning layer," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[37] J. Redmon *et al.*, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 779–788.

[38] D. A. Hudson *et al.*, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6700–6709.

[39] L. H. Li *et al.*, "What does BERT with vision look at?" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 5265–5275.

[40] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[41] Y.-C. Chen *et al.*, "Uniter: Universal image-text representation learning," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 104–120.

[42] J. Liu *et al.*, "Phd: A chatgpt-prompted visual hallucination evaluation dataset," in *CVPR*, 2025.

[43] Y. Li *et al.*, "Evaluating object hallucination in large vision-language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 292–305.

[44] A. Rohrbach *et al.*, "Object hallucination in image captioning," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Oct.-Nov. 2018, pp. 4035–4045.

[45] S. Bai *et al.*, "Qwen3-vl technical report," 2025.

[46] C. Clark *et al.*, "Molmo2: Open weights and data for vision-language models with video understanding and grounding," 2026.

[47] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.