WHAT "NOT" TO DETECT: NEGATION-AWARE VLMS VIA STRUCTURED REASONING AND TOKEN MERGING

Inha Kang[1] **Youngsun Lim** [1] **Seonho Lee** [1] **Jiho Choi** [1] **Junsuk Choe** [2] **Hyunjung Shim**[1*]
[1]KAIST AI   [2]Sogang University
{rkswlsj13, youngsun_ai, glanceyes, jihochoi, kateshim}@kaist.ac.kr
jschoe@sogang.ac.kr

ABSTRACT

State-of-the-art vision-language models (VLMs) suffer from a critical failure in understanding negation, often referred to as affirmative bias. This limitation is particularly severe in described object detection (DOD) tasks. To address this, we propose two primary contributions: (1) a new dataset pipeline and (2) a novel, lightweight adaptation recipe. First, we introduce CoVAND, a dataset constructed with a systematic chain-of-thought (CoT) and VQA-based pipeline to generate high-quality, instance-grounded negation data. Second, we propose NEGTOME, a novel text token merging module that directly tackles the architectural cause of affirmative bias. NEGTOME fundamentally addresses the structural loss of negation cues in tokenization, grouping them with attributes into coherent semantic phrases. It maintains correct polarity at the input level, enabling robust negation understanding even with limited data. For instance, to prevent a model from treating the fragmented tokens `not` and `girl` as simply girl, NEGTOME binds them into a single token whose meaning is correctly distinguished from that of `girl` alone. This module is integrated with a parameter-efficient and strategic LoRA fine-tuning approach. Our method significantly improves performance on challenging negation benchmarks with a lowered false positive rate, boosting NMS-AP by up to +10.8 points on OVDEval and demonstrating generalization to SoTA VLMs. This work marks a crucial step forward in addressing negation understanding for real-world detection applications.

1 INTRODUCTION

Even state-of-the-art Vision-Language Models (VLMs) exhibit a critical failure in understanding negation due to an affirmative bias (Alhamoud et al., 2025). This bias reflects a model's tendency to prioritize nouns while ignoring crucial negation cues. The issue is particularly pronounced in *described object detection* (DOD) (Xie et al., 2023; Schulter et al., 2023; Yao et al., 2024; Dang et al., 2023), a task requiring fine-grained compositional reasoning. As in Figure 1a, this bias causes models to treat phrases like *"person with skateboard"* and *"person **without** skateboard"* as semantically equivalent, leading to identical and incorrect detections. This failure extends to more complex logical structures, such as double negatives (*e.g., "not" + "un-"*). Since humans naturally use negation in natural communication (Sarabi & Blanco, 2016; Morante & Blanco, 2021; Beukeboom et al., 2020; Morante & Sporleder, 2012), failing to handle negation poses a serious barrier to real-world scenarios. This shortcoming can be particularly dangerous in safety-critical domains. For example, in medical imaging, misinterpreting the distinction between *"a tumor that is **not** malignant"* and *"a tumor that is malignant"* can lead to critical misdiagnoses. Therefore, bridging and improving negation understanding is an important step toward building robust VLM-based detection systems.

One key reason for the limited negation capability of VLMs is the lack of negated expressions in existing pre-training datasets. For example, large-scale datasets such as LAION-400M (Schuhmann et al., 2021) contain about 0.08% negation words (Park et al., 2025). Likewise, Flickr30k (Plummer et al., 2015), a widely used captioning dataset, exhibits only 0.04% negation words (Figure 1b). In contrast, negation is much more prevalent in real-world language. For instance, 13.76% of words

---

*Corresponding author

(a) Qualitative Comparison on Challenging Negation Queries.

"person **with** skateboard" vs. "person **without** skateboard"   "**un**peeled banana" vs. "banana that is **not un**peeled"

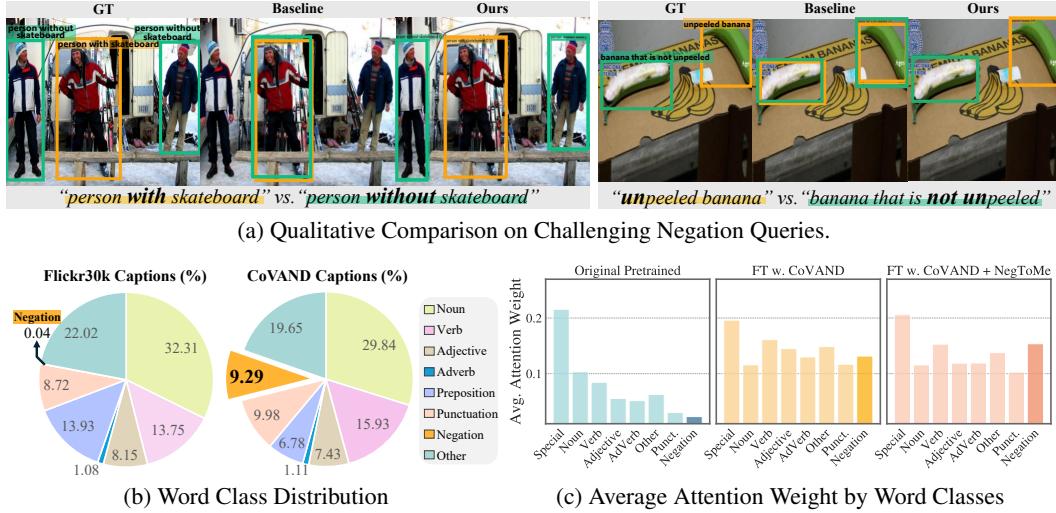(b) Word Class Distribution   (c) Average Attention Weight by Word Classes

Figure 1: **Challenges with Negation Expressions.** (a) Standard VLMs exhibit an affirmative bias, failing to distinguish contradictory negation queries. This issue stems from two causes: (b) the scarcity of negation words in standard datasets and (c) the model's tendency to assign low attention to negation cues. Our solutions, COVAND and the NEGTOME, directly address both problems.

in scientific papers (Szarvas et al., 2008) and 22.23% of words in Conan Doyle's stories involve negation (Morante & Daelemans, 2012). This imbalance results in VLMs that are poorly equipped to learn or attend to negation semantics.

To mitigate this limitation, we introduce a **c**hain-**o**f-thought with **V**QA **a**lignment for **n**egation **d**etection dataset (COVAND). It is a negation-focused training dataset constructed via chain-of-thought (CoT) reasoning and VQA-based caption alignment. To construct COVAND, we first extract both present and absent attributes from object regions. For each region, we then generate matched positive and negative captions using a CoT approach, followed by semantic verification using a VQA module. This process ensures each caption precisely reflects the presence or absence of key attributes, resulting in high-quality negation data pairs. As a result, our dataset provides a rich resource with 9.29% of negation words, a frequency 100× higher than that of typical datasets.

In addition to data-related factors, we observe that negation tokens receive notably lower attention weights, suggesting that current VLM detectors architecturally ignore or undervalue negation cues, as shown in Figure 1c. To counteract the low attention given to negation cues, the core of our method is NEGTOME, our novel text token merging module. It is designed to solve a key problem where standard tokenization often fragments phrases, separating negation cues (e.g., "not") from the attributes they modify (e.g., "lying"). NEGTOME addresses this by first merging these fragmented tokens into a single, coherent phrase. Through this binding, the negated concept of "not lying" can be learned as semantically distinct from "lying". This step strengthens the role of the attribute by ensuring it is always interpreted within its negated context. Crucially, this merged representation is enhanced with a negation-aware boost, explicitly amplifying the negated signal to ensure its polarity is preserved for downstream fusion. To our knowledge, this is the first work to employ a boosted token merging strategy for preserving semantic polarity in VLM-based detection.

To ensure the model effectively uses this enhanced text representation, we combine NEGTOME with a highly targeted application of Low-Rank Adaptation (LoRA). Our layer-wise attention analysis revealed that the negation signal dissipates before reaching the final decision-making blocks. Therefore, we apply LoRA to the deep cross-attention layers, the core of multimodal compositional understanding (Laurençon et al., 2024; Hertz et al., 2022). Together, this strategy modifies less than 0.1% of the model's parameters yet achieves a significant improvement in negation comprehension.

Our approach achieves state-of-the-art performance with 6.6 mAP on D³ dataset, with 7.2 mAP improvement specifically on the challenging absence subset. In particular, our method not only increases the NMS-AP metric by 10.8 mAP but also reduces the false positive rate by 19.1%, demonstrating its enhanced ability to distinguish between contradictory queries. Importantly, these results are consistently observed across multiple distinct evaluation datasets, despite the model being
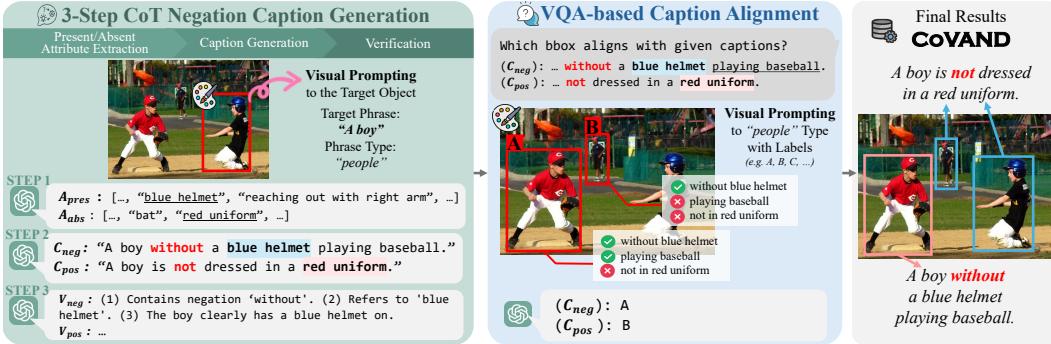
Figure 2: **Dataset Generation Pipeline of the CoVAND.** Our method first generates negation-focused captions for visually prompted regions using a three-step CoT process, then aligns each caption with the correct bounding box via VQA-based reasoning to ensure semantic correspondence.

trained solely on CoVAND. This highlights the strength of our approach and its superior generalization capability to unseen data and negation patterns.

Our work represents an initial yet substantial step toward robust negation understanding with the following key contributions:

- Our work presents CoVAND, a systematically generated dataset focusing on negation, to bridge a critical gap within existing multimodal benchmarks.

- We propose a novel adaptation recipe with NEGTOME, our text token merging module that introduces a negation-aware boost to preserve semantic polarity.

- We achieve consistent gains across benchmarks, including +7.2 mAP on D^3 absence subset and +10.8 mAP on the NMS-AP metric in OVDEval's negation subset, demonstrating effective generalization to real-world negation scenarios.

2 CoVAND: DATASET GENERATION

To address the scarcity of negation data, we present CoVAND, a region-grounded negation dataset constructed through a multi-stage pipeline. As shown in Figure 2, the curation process consists of CoT caption generation followed by VQA-based alignment. This pipeline generates new high-quality captions that cover not only existence but also diverse attribute-based negations. In this way, CoVAND provides fine-grained, compositional supervision that trains detectors more robustly than only injecting templated or caption-level negations (Alhamoud et al., 2025; Park et al., 2025).

2.1 VISUAL PROMPTING WITH BOUNDING BOXES

Before caption generation, we apply visual prompting (Cai et al., 2024) to overlay a marker on the image. The marker specifies the region to describe and directs the CoT model's attention to that area. We apply this technique to bounding boxes in the Flickr30k Entities dataset (Plummer et al., 2015). For each image, we randomly choose two boxes linked to meaningful objects and exclude any box that spans a large background area to avoid ambiguity. Each selected region is then highlighted with a red bounding box and serves as an input image for region-grounded caption generation.

2.2 THREE-STEP CHAIN-OF-THOUGHT CAPTION GENERATION

We generate region-grounded paired negation captions through a three-step CoT process using GPT-4o (Hurst et al., 2024). We provide an explicit sequence that ensures consistent quality, rather than leaving it to the model's decision. The design follows the multi-step reasoning strategy of LLMs, where a complex visual query is split into ordered subtasks that improve factual accuracy and transparency. The input prompt for caption generation shows the image with a red bounding box, a target phrase such as "a boy" in "person" type. These cues fix the subject within the highlighted area and guide each reasoning step. The three steps are detailed below.

Step 1: Present and Absent Attribute Extraction. For each visually prompted region, we extract two sets of attributes: (1) *Present Attributes* (A_{pres}), consisting of attributes visibly present within the bounding box (e.g., colors, actions, relationships, actions, etc.), and (2) *Absent Attributes* (A_{abs}), representing relevant but missing attributes that could reasonably be expected. This rich attribute pool is the key novelty that lets our pipeline create attribute-level negations, which are far beyond the object-level attributes used in prior approaches (Alhamoud et al., 2025).

Step 2: Negative and Positive Caption Generation. We generate two types of paired captions using the extracted attributes:

- *Negative Caption* (C_{neg}): Incorrectly describes an attribute in A_{pres} as absent (e.g., *"A man without a hat"* when "hat" $\in A_{pres}$).
- *Positive Caption* (C_{pos}): Correctly describes an attribute in A_{abs} as absent (e.g., *"A woman without a red hoodie"* when "red hoodie" $\in A_{abs}$).

Each caption includes negation cues such as "no", "not", "never", "without", the prefix "un-", or the contraction "n't". The cue list is open to keep language natural and diverse.

Step 3: Verification. To ensure semantic consistency, we verify that C_{pos} accurately describes the region while C_{neg} contradicts it by asking GPT-4o. We also check whether generated captions contain negation words and attributes from step 1. If the pair fails on the test, it discards invalid captions and repeats caption generation until a valid pair appears or the retry limit is reached. This iterative guard preserves semantic integrity and keeps the quality of the overall dataset.

2.3 VQA-BASED CAPTION ALIGNMENT

The CoT stage produces a positive caption C_{pos} and a negative caption C_{neg} for each randomly chosen target box. However, label noise may still occur since another object of the same phrase type can also fit the captions. In Figure 2, for example, a person marked with "A" in the image could satisfy C_{neg}, even though it is not the designated target, which causes label noise. To eliminate this ambiguity, we add a dedicated region-level VQA alignment step.

First, we draw alphabetical labels on every box that shares the phrase type of the target. The target box stays unlabelled because it has already passed the in-context verification step. To determine the final alignment, we ask a VQA model two separate questions: "Which labelled box aligns with C_{pos}/C_{neg}?". Then, the VQA model simply answers with overlayed letters on the input images. While prior work used VQA for coarse, image-level validation (Park et al., 2025), their approach fails to resolve which specific instance a caption refers to. Our region-level alignment stage solves this ambiguity by requiring the VQA model to match each caption to a specific, visually-labeled bounding box, thereby delivering a more region-level ground truth.

Through this multi-stage process combining CoT reasoning and VQA alignment, CoVAND provides rich training signals for negation understanding. We generate 91,110 captions with 23,876 images. In particular, our dataset exhibits approximately 9.29% negation word frequency, significantly higher than existing datasets like Flickr30k (0.04%). Detailed examples in Appendix A.

3 FINE-TUNING WITH NEGATION-SENSITIVE TEXT TOKEN MERGING

Our method addresses the two root causes of negation blindness: token fragmentation and low attention on negation cues. We propose a lightweight adaptation recipe that integrates our novel text token merging module, NEGTOME, with a targeted application of LoRA as in Figure 3.

3.1 NEGATION LORA ADAPTER

We apply LoRA following (Hu et al., 2021) with two key enhancements for vision-language fusion. Given frozen base weights $W_q, W_v \in \mathbb{R}^{d \times d}$ in cross-attention layers, we inject parallel adapters with an activation layer. Let $\sigma(\cdot)$ denote ReLU (Agarap, 2018) and let $A_q, A_v \in \mathbb{R}^{r \times d}$ and $B_q, B_v \in \mathbb{R}^{d \times r}$ be the trainable low-rank matrices. For an input $x \in \mathbb{R}^d$ we obtain

$$q = W_q x + \alpha B_q \sigma(A_q x), \quad v = W_v x + \alpha B_v \sigma(A_v x), \tag{1}$$

where $W_q, W_v \in \mathbb{R}^{d \times d}$ are the frozen base weights and α scales the LoRA update.
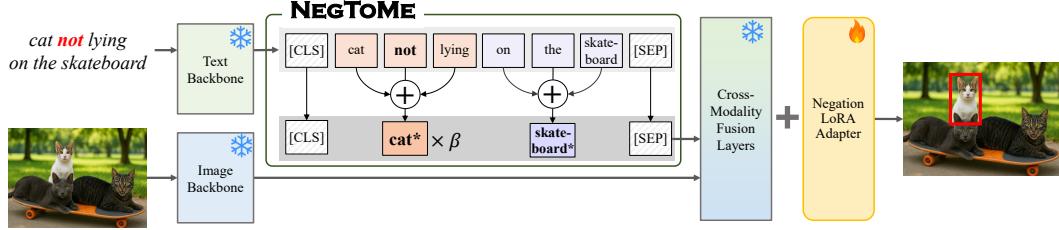
Figure 3: **Overview of Training Pipeline.** The input image and captions of CoVAND are encoded by frozen backbones. NegToMe assigns higher importance to negation cues in the text, and the LoRA adapter enables accurate localization of objects described by negated queries.

3.2 NegToMe: Semantic Text Token Merging for Negation Understanding

Motivation. While fine-tuning with negation-rich data can partially alleviate affirmative bias, it does not address a more fundamental flaw embedded in the model's tokenization process. Standard tokenizers inherently fragment phrases, separating negation cues (e.g., "not") from the words they modify (e.g., "lying"). This structural separation effectively causes the model to treat the phrase "not lying" as semantically equivalent to "lying", as the attention weight of the isolated negation tends to be ignored. To rectify this intrinsic information loss, we introduce NegToMe. It moves beyond data-level fixes to structurally ensure that a negated concept like "cat not lying" is represented as a single semantic unit, fundamentally distinct from {"cat", "not", "lying"}.

Text Token Merging. The caption is first split into sub-tokens $\mathcal{T} = \{t_1, \ldots, t_n\}$ by a standard tokenizer. To merge the tokens, an off-the-shelf parser then groups these tokens into disjoint phrase sets $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_m\}$ where $m < n$. For every phrase $\mathcal{P}_i \subseteq \mathcal{T}$, we compute one representative embedding by taking the softmax–weighted average of the sub-token vectors inside the phrase and replacing the original vectors with this average.

Negation-aware Boost. After merging, let $\mathcal{P}_{\text{neg}}$ be the phrase containing a cue (not, no, without, un-, etc.), and $\mathcal{I}_{\text{neg}} = \{ j \mid t_j \in \mathcal{P}_{\text{neg}} \}$ its index set. We assign a larger weight to the negation cue:

$$\bar{t}_{\text{neg}} = \frac{\sum_{j \in \mathcal{I}_{\text{neg}}} \gamma_j \, t_j}{\sum_{j \in \mathcal{I}_{\text{neg}}} \gamma_j}, \qquad \gamma_j = \begin{cases} \beta & \text{if } t_j \text{ is the negation cue,} \\ 1 & \text{otherwise,} \end{cases} \qquad \beta > 1. \qquad (2)$$

The negation boosting factor β amplifies the cue so that the merged embedding explicitly retains the negated meaning, improving polarity reasoning without increasing sequence length.

Effect of Negation Boost on Representations. Suppose the encoder maps a caption of n sub-tokens to vectors $h_1, \ldots, h_n \in \mathbb{R}^d$. We write h_c for the vector of the negation cue (e.g. "*not*") and h_p for the vector of the predicate it modifies (e.g. "*moving*"). With vanilla mean pooling, the sentence embedding is $\bar{h} = \frac{1}{n} \sum_{i=1}^n h_i$, so the cue contributes only $s_{\text{single}} = \langle v, h_c \rangle / n$ to any linear probe $v \in \mathbb{R}^d$. After applying NegToMe, the merged representation of the negated phrase becomes $h_{\text{neg}} = \frac{\beta \, h_c + h_p}{\beta + 1}$ and the pooled vector gives $s_{\text{merge}} \geq \frac{\beta}{\beta+1} \langle v, h_c \rangle / m$, Hence

$$\frac{s_{\text{merge}}}{s_{\text{single}}} \geq \frac{\beta}{\beta + 1} \cdot \frac{n}{m}, \qquad 1 \leq m < n, \qquad (3)$$

so the cue's influence is amplified by at least the factor $\frac{\beta}{\beta+1} \cdot \frac{n}{m}$. This gain aligns with the larger attention weights observed in Figure 1c and Figure S15, and experimentally show higher mAP.

4 Experiments

4.1 Experimental Setups

Datasets. DOD requires resolving compositional descriptions as in Figure 4a. To rigorously assess our model's ability to overcome the affirmative bias inherent in VLMs, we select two benchmarks specifically designed to challenge negation understanding. We evaluate our method on two challenging DOD benchmarks for negation detection in VLMs. *Described Object Detection* (D^3) (Xie et al.,

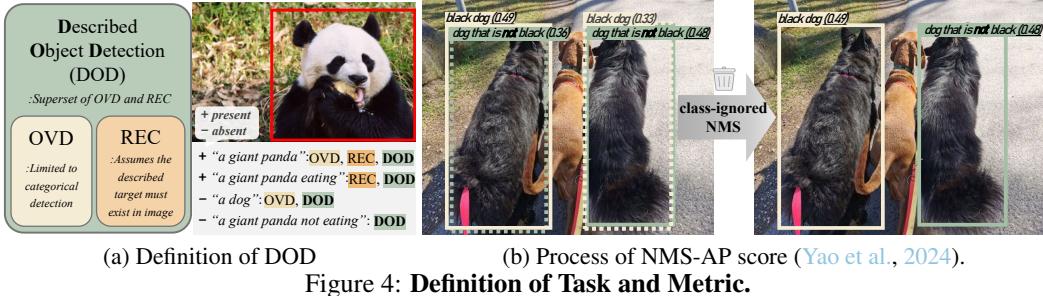(a) Definition of DOD   (b) Process of NMS-AP score (Yao et al., 2024).

Figure 4: **Definition of Task and Metric.**

2023) introduces three evaluation protocols. *Pres* is a subset of 316 presence descriptions, *ABS* is 106 absence descriptions, and *Full* is an evaluation across all 422 descriptions. For *OVDEval Negation Subset* (Yao et al., 2024), we report both standard AP and the NMS-AP. The standard AP score can be misleadingly inflated when a model, confused by fragmented tokens, predicts overlapping boxes for contradictory pairs like *"black dog"* and *"dog that is **not** black"*. In contrast, NMS-AP (Yao et al., 2024) applies stricter filtering by removing overlapping predictions on contradictory pairs with IoU>0.5, effectively penalizing affirmative bias and accurately measuring negation understanding (Figure 4b). Additionally, we employ a practical yet challenging evaluation by performing class-ignored NMS separately after predicting each caption individually. (see the Appendix D.1.)

Implementation Details. We implement parameter-efficient fine-tuning through LoRA (Hu et al., 2021) applied to the deep cross-attention layers in the vision-language fusion module with $r = 4$. VLM-based detectors are trained for $5,000$ iterations with a batch size of 24 for the Grounding DINO model, and $6,000$ iterations with a batch size of 4 for the APE-Ti model. Training is conducted on two NVIDIA A6000 GPUs with mixed precision with a learning rate of 5×10^{-4}. Qwen-2.5-VL (Bai et al., 2025) is trained for 1 epoch batch size of 32 with a learning rate of 5×10^{-5}. All models are only trained with the COVAND dataset using the AdamW optimizer (Loshchilov & Hutter, 2017), freezing all backbone parameters except the LoRA layers. For NEGTOME, we use spaCy for the parser and set the negation boost factor $\beta = 2.0$. More details in the Appendix B.

4.2 EXPERIMENTAL RESULTS

Quantitative Results. As shown in Table 1, even powerful Multimodal Large Language Models (MLLMs) struggle with the D^3 benchmark. SoTA models like SPHINX-7B (Lin et al., 2023) and Qwen-2.5-VL-3B (Bai et al., 2025) achieve low performance on the full set (10.6 and 18.6 mAP, respectively), and their slow inference makes them impractical for many detection scenarios. In contrast, our lightweight adaptation recipe significantly boosts the performance of strong detector baselines. When applied to Grounding-DINO, our method improves the overall mAP by +6.6 points, with a notable gain of **+7.2 mAP** on the challenging absence subset. This performance gain is direct evidence of a more robust understanding of semantic polarity. Baseline models often generate

Table 1: **Evaluation on the D^3 benchmarks.** Descriptions categorized by length; *S* for 1-3, *M* for 4-6, *L* for 7-9, and *XL* for 10+ words. *Pres* refers to present and *Abs* refers to absence subset.

Method	Architecture			D^3 (default)			D^3 (by length of texts)			
	Backbone	Text Encoder	Detection Head	Full	Pres	Abs	S	M	L	XL
OFA-L	ResNet-101+ViT	BART	Seq2Seq	4.2	4.1	4.6	4.9	5.4	3.0	2.1
OWL-ViT-L	ViT-L	CLIP	OWL-ViT	9.6	10.7	6.4	20.7	9.4	6.0	5.3
SPHINX-7B	CLIP,DINO-v2, Q-Former	LLaMA-2	-	10.6	11.4	7.9	16.8	13.8	5.6	3.1
OFA-DOD	ResNet-101+ViT	BART	Seq2Seq	21.6	23.7	15.4	23.6	22.6	20.5	18.4
GLIP-T				19.1	18.3	21.5	22.4	22.0	16.6	10.6
+ GEN	Swin-T	BERT	DyHead	21.4	20.6	23.7	28.1	24.5	17.4	11.5
+ W2S				26.0	25.6	27.1	-	-	-	-
FIBER-B				22.7	21.5	26.0	30.1	25.9	17.9	13.1
+ GEN	Swin-B	RoBERTa-B	DyHead	26.0	25.2	28.1	35.5	29.7	20.5	14.2
+ W2S				26.5	26.0	27.7	-	-	-	-
G-DINO-B				20.7	20.1	22.5	22.6	22.5	18.9	16.5
+ Ours	Swin-B	BERT	DINO	27.3	26.4	29.7	29.9	29.5	25.2	21.3
(↑ Δ)				(+6.6)	(+6.3)	(+7.2)	(+7.3)	(+7.0)	(+6.3)	(+4.8)
APE-Ti				29.1	29.9	26.9	31.1	31.9	27.4	21.4
+ Ours	ViT-Ti	CLIP	DETA	32.5	32.9	31.5	33.2	35.3	31.3	25.4
(↑ Δ)				(+3.4)	(+3.0)	(+4.6)	(+2.1)	(+3.4)	(+3.9)	(+4.0)
Qwen-2.5-VL-3B				18.6	18.5	19.2	18.2	20.7	17.0	16.0
+ Ours	ViT-H	Qwen-2.5	–	22.2	22.8	20.6	19.8	25.8	20.2	17.8
(↑ Δ)				(+3.6)	(+4.3)	(+1.4)	(+1.6)	(+5.1)	(+3.2)	(+1.8)

false positives because they fail to distinguish between conceptually opposite phrases like "*with a hat*" and "*without a hat*". As a specific absence scenario, when prompted with "*a person without a hat*" in an image where everyone is wearing one, they would incorrectly detect a person. Our tokenizer modification, NEGTOME, resolves this by forcing the model to process the negated phrase as a single semantic unit with distinct polarity, enabling it to correctly reject such invalid instances. Similarly, on APE-Ti, we achieve a +4.6 mAP improvement on the absence subset, demonstrating an enhanced ability to reject non-existent objects. Notably, these gains are comparable to computationally expensive, large-scale fine-tuning methods (Zhao et al., 2024a; Park et al., 2024) while updating less than 0.1% of the model's parameters only with our COVAND dataset. The improvements are also consistent across all description lengths, validating the robustness of our approach. Furthermore, preliminary experiments demonstrate the generalizability of our method to MLLMs, with an improvement of +3.6 mAP on Qwen-2.5-VL-3B.

Even powerful SoTA MLLMs struggle on the challenging OVDEval-Negation subset, demonstrating that simply applying a large-scale model is not a sufficient solution for negation. Notably, as shown in Table 2, the powerful Qwen-2.5-VL-7B underperforms the much smaller Grounding-DINO baseline, highlighting the difficulty of the task. In contrast, our lightweight adaptation recipe yields significant performance gains across all tested architectures, particularly on the stricter NMS-AP metric. Our method boosts the Grounding-DINO by a substantial **+10.8** mAP in NMS-AP and improves the Qwen-2.5-VL-3B by +7.3 in mAP and +3.8 in NMS-AP. For the MLLM, the substantial AP gain is significant because it enhances both negation reasoning and foundational localization, a typical weakness of such models. Further results, including a detailed comparison with two-stage post-hoc VQA with MLLM and a full evaluation across all OVDEval subsets, are available in Appendix C and D.

Table 2: **Results on OVDEval-Negation**. [†] means reproduced AP.

	AP	NMS-AP
G-DINO-B[†]	54.0	36.8
+ Ours	57.2	47.6
(↑ Δ)	(+3.2)	(+10.8)
APE-Ti	50.5	32.3
+ Ours	54.1	33.5
(↑ Δ)	(+3.6)	(+1.2)
Qwen-2.5-VL-7B	37.8	35.9
Qwen-2.5-VL-3B	34.6	31.3
+ Ours	41.9	35.1
(↑ Δ)	(+7.3)	(+3.8)

Dataset	# Images	# Captions
COVAND-S	8k	30.6k
COVAND-M	16k	60.4k
COVAND-L	24k	91.1k

(a) COVAND splits.　　　(b) Dataset Scalability. NMS-AP (left) and FPR (right).
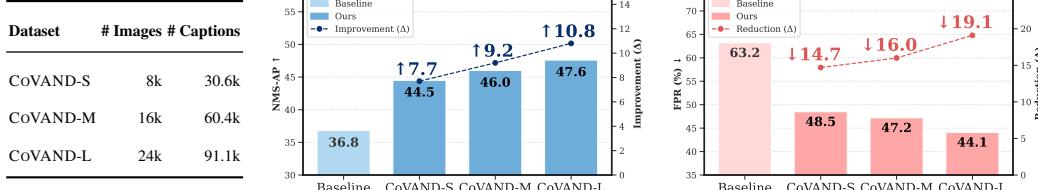
Figure 5: **Dataset Statistics and Performance Scaling.** (a) Statistics for our three COVAND splits. (b) Bar plots with **blue** refer to NMS-AP and **pink** refer to FPR (lower is better).

Dataset Scalability. Figure 5 presents our scalability analysis of the dataset on the OVDEval-Negation subset. We observe a consistent improvement as we scale the COVAND dataset from small to large. Specifically, NMS-AP improves from 44.5 to 47.6, while the FPR decreases from 48.5% to 44.1%, which is a total reduction of **19.1** points from the baseline. This trend of simultaneously improving NMS-AP, a metric that penalizes contradictory predictions, while lowering FPR, which measures the failure to reject absent objects, shows the effectiveness of our approach.

Qualitative Results. Figure 6 presents qualitative results from the OVDEval dataset comparing our fine-tuned Grounding DINO model against the baseline. The baseline model often exhibits a strong affirmative bias, frequently collapsing contradictory captions into the same prediction. Our model, however, successfully handles these complexities across various patterns. For instance, it accurately identifies the "*cow without looking at the camera*" and the "*horse that is not urinating*", proving it can ground negation in complex contexts. Moreover, for "*banana that is not unpeeled*", it correctly identifies the peeled banana by resolving the "not" + "un-" double negative as in Figure 1a. Our model sometimes fails to detect every target instance, for example "*pizza that is not complete*", its predictions are a marked improvement over the baseline, which provides completely unreliable detections for both queries. Together, these examples show that our method achieves a more compositional understanding of negation. Further qualitative results on OVDEval and D^3 are presented in Figure S20–S21 and Figure S22–S24, respectively.
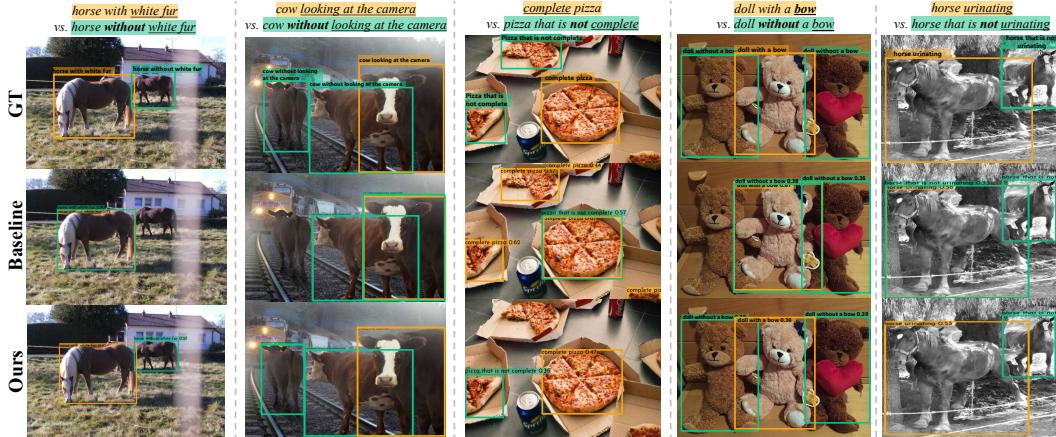
Figure 6: **Qualitative Comparison on the OVDEval Negation Subset.** Our model correctly distinguishes the polarity of contradictory caption pairs, overcoming the baseline's affirmative bias.

Table 3: **Ablation Study.** Best in blue and worst in red. LoRA adapters are inserted at three fusion-block depths: `shallow` (blocks 0–2), `strided` (1, 3, 5), and `deep` (3–5).

Training Data	Settings LoRA Placement	NEGTOME	β	OVDEval (Negation Subset) AP	NMS-AP	AR	NMS-AR	↓FPR	D^3 Full	Pres	Abs	↓FPR
Pretrained Weight				54.0	36.8	20.5	14.7	63.2	20.7	20.1	22.5	67.2
Flickr30k	shallow	✗	–	55.9	38.5	21.7	15.2	61.3	18.4	18.2	23.0	66.5
Flickr30k	strided	✗	–	54.8	36.5	20.5	14.1	62.6	20.9	19.9	24.0	68.2
Flickr30k	deep	✗	–	53.7	31.8	20.7	12.8	59.9	22.0	21.0	24.8	67.8
COVAND-S	shallow	✗	–	46.8	31.5	21.9	14.8	56.0	18.5	17.6	21.0	63.9
COVAND-S	strided	✗	–	52.8	43.9	20.0	17.1	49.0	20.1	19.2	22.9	63.4
COVAND-S	deep	✗	–	55.4	41.8	21.4	18.0	48.6	24.2	23.0	27.0	64.0
COVAND-S	deep	✔	1.0	57.8	43.8	24.0	19.6	50.8	25.7	25.1	27.3	63.7
COVAND-S	deep	✔	2.0	58.7	44.5	24.1	19.2	48.5	26.2	25.4	28.2	63.3

4.3 ABLATION STUDY

Our ablation study, summarized in Table 3, reveals the impact of each component, with attention diagnostics in Figure S15 in the Appendix providing a clear mechanism for the improvements. Placing LoRA adapters in the `deep` fusion blocks consistently outperforms `shallow`. This is because `deep` placement maintains elevated attention on negation tokens in the later blocks where decisions are formed, whereas the effect of `shallow` placement dissipates too early. Furthermore, training with COVAND dataset yields substantial gains over generic captions, demonstrating its value for both accuracy and generalization. Finally, adding NEGTOME with its negation boost factor provides large gains, such as a **+2.7** improvement in NMS-AP. This trend is mirrored on the D^3 benchmark. While using our COVAND dataset alone yields a +2.2 mAP improvement over the baseline, NEGTOME adds a further +2.0 mAP on top. This near-equal contribution highlights that our token merging strategy is as impactful as the dataset itself. The attention analysis further confirms that NEGTOME directly causes this improvement by increasing attention to the negated phrase.

4.4 ZERO-SHOT DOWNSTREAM EVALUATION OF SEMANTIC COMPREHENSION.

To verify our method achieves a semantic understanding of negation that generalizes beyond detection, we evaluate it on the NegBench COCO subset of Multiple Choice Question (MCQ) benchmark (Alhamoud et al., 2025). This task requires the model to select the most accurate caption for an image from four options. These

Table 4: **Results on the NegBench Multiple Choice Question (MCQ) benchmark.**

Model	Overall Acc.	Positive	Negative	Hybrid
CLIP-OpenAI	16.27 %	—	—	—
NegCLIP	10.21 %	—	—	—
G-DINO-B	21.69 %	27.36 %	13.37 %	23.71 %
+ Ours	**32.55 %**	**46.85 %**	**23.37 %**	**26.64 %**
(↑ Δ)	(+10.86)	(+19.49)	(+10.00)	(+2.93)

options include three subsets: 'Positive' correctly affirming present objects (e.g., "*A and B*"), 'Negative' correctly negating absent ones (e.g., "*not B*"), and 'Hybrid' that combine both types (e.g., "*A but not B*"). In a zero-shot setting, we select the caption that produces the highest max-logit score when grounded in the image. As shown in Table 4, our method improves accuracy over the baseline with a +10.86% improvement. This result provides strong evidence that our approach enhances a robust understanding of negation. We present qualitative examples in Figure 7 and in Appendix F.

Figure 7: Qualitative Comparison on the NegBench MCQ Benchmark. Captions with green checkmark ✅ is **GT**, pink refer to **Baseline**, and blue refer to **Ours**.

5 RELATED WORK

5.1 OBJECT DETECTION

OVD extends classical detectors to arbitrary text labels (Zareian et al., 2021; Yao et al., 2022; Kim et al., 2023). Methods such as GLIP (Li et al., 2022), and APE (Shen et al., 2024) fuse language either in the detection head, in the backbone, or in a task-general prompt module, and achieve strong zero-shot performance. REC adds compositional phrases. Grounding DINO (Liu et al., 2024) proposes DETR-style decoders that localize the described object without category supervision. Despite this progress, REC models still assume the target exists and therefore struggle to reject absent or negated descriptions. DOD (Xie et al., 2023) generalizes OVD and REC by requiring the detector to decide both existence and location. Benchmarks such as D^3 and OVDEval (Yao et al., 2024) reveal a low in accuracy on absence or negation subsets. It confirms that current VLMs often have an affirmative bias on negation cues. MLLM (Lin et al., 2023; Bai et al., 2025) have recently been applied to DOD, but their accuracy fails to surpass that of VLM-based detectors, their performance on negation remains low, and their inference speed is incompatible with real-time detection scenarios.

5.2 NEGATION UNDERSTANDING IN VISION-LANGUAGE MODELS

CLIP-based studies such as NegBench (Alhamoud et al., 2025) reveal the affirmative bais that state-of-the-art VLMs often treat *"dog"* and *"not dog"* identically; subsequent fixes like Negation-CLIP (Park et al., 2025) simply augment pre-training with template-level negation pairs and thus miss context-dependent or region-grounded cases. We instead build a fine-grained dataset with CoT reasoning and VQA alignment, producing positive and negative caption pairs that are grounded to target boxes, and show that this richer supervision transfers to multiple architectures beyond CLIP.

5.3 TEXT TOKEN-LEVEL MERGING

Token Merging (ToMe) (Bolya et al., 2022) merges similar image tokens to accelerate inference without sacrificing accuracy. ToMe is extended to diffusion and grounding models, where token merging based on semantic phrase is introduced to mitigate the loss of modifier information (Hu et al., 2024; Li et al., 2024). In the context of OVD, there have been attempts to merge image tokens (Su et al., 2024; Norouzi et al., 2024), but the merging of text tokens has been unexplored. Previous studies on text token merging have primarily focused on diffusion models, particularly in text-to-image generation (Hu et al., 2024). In this work, we are the first to explore text token merging in detection models and empirically demonstrate its feasibility and effectiveness.

6 CONCLUSION

This work presents a comprehensive solution to the affirmative bias that hinders negation understanding in VLMs by addressing its two root causes. To resolve data scarcity, we introduce CO-VAND, a systematic pipeline using CoT reasoning and VQA-based alignment to generate high-quality, instance-grounded negation data. To counteract the model's architectural tendency to ignore negation cues, we propose NEGTOME, a novel module that, to our knowledge, is the first to use a negation-aware boost to preserve semantic polarity in detection tasks. Our parameter-efficient recipe integrates these contributions to achieve substantial gains on challenging negation benchmarks and demonstrate strong generalization across VLM-based detectors and MLLMs, marking a significant step towards VLMs that can understand not only what is present, but also what is absent.

BIBLIOGRAPHY

Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. *arXiv preprint arXiv:2501.09425*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Camiel J Beukeboom, Christian Burgers, Zsolt P Szabó, Slavica Cvejic, Jan-Erik M Lönnqvist, and Kasper Welbers. The negation bias in stereotype maintenance: A replication in five languages. *Journal of Language and Social Psychology*, 39(2):219–236, 2020.

Franziska Boenisch, Kamil Deja, Adam Dziedzic, et al. Precise parameter localization for textual generation in diffusion models. *arXiv preprint arXiv:2502.09935*, 2025.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Making large multimodal models understand arbitrary visual prompts. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

Ronghao Dang, Jiangyan Feng, Haodong Zhang, Chongjian Ge, Lin Song, Lijun Gong, Chengju Liu, Qijun Chen, Feng Zhu, Rui Zhao, et al. Instructdet: Diversifying referring object detection with generalized instructions. *arXiv preprint arXiv:2310.05136*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35:32942–32956, 2022.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

Chongyang Gao, Kezhen Chen, Jinmeng Rao, Baochen Sun, Ruibo Liu, Daiyi Peng, Yawen Zhang, Xiaoyuan Guo, Jie Yang, and VS Subrahmanian. Higher layers need more lora experts. *arXiv preprint arXiv:2402.08562*, 2024.

Chongyang Gao, Kezhen Chen, Jinmeng Rao, Ruibo Liu, Baochen Sun, Yawen Zhang, Daiyi Peng, Xiaoyuan Guo, and Vs Subrahmanian. MoLA: MoE LoRA with layer-wise expert allocation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5097–5112, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. URL https://aclanthology.org/2025.findings-naacl.284/.

Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2918–2928, 2021.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Edward Hu et al. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Taihang Hu, Linxuan Li, Joost van de Weijer, Hongcheng Gao, Fahad Shahbaz Khan, Jian Yang, Ming-Ming Cheng, Kai Wang, and Yaxing Wang. Token merging for training-free semantic binding in text-to-image synthesis. *Advances in Neural Information Processing Systems*, 37: 137646–137672, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1780–1790, 2021.

Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11144–11154, 2023.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024.

Liunian Li, Zi-Yi Dou, Nanyun Peng, and Kai-Wei Chang. Desco: Learning object recognition with rich language descriptions. *Advances in Neural Information Processing Systems*, 36:37511–37526, 2023.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10965–10975, 2022.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *arXiv preprint arXiv:2407.02392*, 2024.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Roser Morante and Eduardo Blanco. Recent advances in processing negation. *Natural Language Engineering*, 27(2):121–130, 2021.

Roser Morante and Walter Daelemans. Conandoyle-neg: Annotation of negation in conan doyle stories. In *Proceedings of the eighth international conference on language resources and evaluation, istanbul*, pp. 1563–1568. Citeseer, 2012.

Roser Morante and Caroline Sporleder. Modality and negation: An introduction to the special issue. *Computational linguistics*, 38(2):223–260, 2012.

Narges Norouzi, Svetlana Orlova, Daan De Geus, and Gijs Dubbelman. Algm: Adaptive local-then-global token merging for efficient semantic segmentation with plain vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15773–15782, 2024.

Junsung Park, Jungbeom Lee, Jongyoon Song, Sangwon Yu, Dahuin Jung, and Sungroh Yoon. Know" no" better: A data-driven approach for enhancing negation awareness in clip. *arXiv preprint arXiv:2501.10913*, 2025.

Kwanyong Park, Kuniaki Saito, and Donghyun Kim. Weak-to-strong compositional learning from generative models for language-based object detection. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2024.

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 *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Zahra Sarabi and Eduardo Blanco. Understanding negation in positive terms using syntactic dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1108–1118, 2016.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Samuel Schulter, Vijay Kumar B G, Yumin Suh, Konstantinos M. Dafnis, Zhixing Zhang, Shiyu Zhao, and Dimitris Metaxas. Omnilabel: A challenging benchmark for language-based object detection. In *ICCV*, 2023.

Dominykas Seputis, Serghei Mihailov, Soham Chatterjee, and Zehao Xiao. Multi-modal adapter for vision-language models. *arXiv preprint arXiv:2409.02958*, 2024.

Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. Aligning and prompting everything all at once for universal visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13193–13203, 2024.

Wei Su, Peihan Miao, Huanzhang Dou, and Xi Li. Scanformer: Referring expression comprehension by iteratively scanning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13449–13458, 2024.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the workshop on current trends in biomedical natural language processing*, pp. 38–45, 2008.

Chi Xie, Zhao Zhang, Yixuan Wu, Feng Zhu, Rui Zhao, and Shuang Liang. Described object detection: Liberating object detection with flexible expressions. *Advances in Neural Information Processing Systems*, 36:79095–79107, 2023.

Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022.

Yiyang Yao, Peng Liu, Tiancheng Zhao, Qianqian Zhang, Jiajia Liao, Chunxin Fang, Kyusong Lee, and Qing Wang. How to evaluate the generalization of detection? a benchmark for comprehensive open-vocabulary detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6630–6638, 2024.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14393–14402, 2021.

Shiyu Zhao, Long Zhao, Yumin Suh, Dimitris N Metaxas, Manmohan Chandraker, Samuel Schulter, et al. Generating enhanced negatives for training language-based object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13592–13602, 2024a.

Tiancheng Zhao, Peng Liu, and Kyusong Lee. Omdet: Large-scale vision-language multi-dataset pre-training with multimodal detection network. *IET Computer Vision*, 18(5):626–639, 2024b.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Wei Li Zuwei Long. Open grounding dino:the third party implementation of the paper grounding dino. https://github.com/longzw1997/Open-GroundingDino, 2023.

SUPPLEMENTARY MATERIALS

We provide supplementary materials in the following order:

- Section A: **CoVAND details** on our negation-focused dataset generation process, including the three-step Chain-of-Thought prompt design and VQA-based caption alignment that ensures precise region-caption correspondence.
- Section B: **Implementation Details** presents architectural specifications and analysis.
- Section C: **Comparison with Post-hoc VQA Methods** analyzes a two-stage, post-hoc VQA approach, comparing its accuracy-latency trade-offs to our single-stage method.
- Section D: **Evaluation on Full OVDEval Subsets** provides results across all OVDEval subsets, demonstrating our model's robust generalization beyond negation-specific tasks.
- Section E: **Analysis on RPN-based Detector** examines why region proposal networks struggle with negation expressions compared to our DETR-based approach.
- Section F: **Detailed Analysis on Zero-shot NegBench Downstream tasks** give information of each subset and error types based on qualification examples.
- Section G: **Qualitative Results** shows visual examples that illustrate our model's improved negation handling capabilities, highlighting reduced false positives and better attribute discrimination under negation.
- Section H: **Author Statements** contains LLM usage, ethics, and reproducibility statement.

A DETAILS ON COVAND

A.1 PROMPT FOR THREE-STEP COT CAPTION GENERATION

We employ a systematic three-step CoT reasoning approach using GPT-4o (Hurst et al., 2024) to generate high-quality negation-focused captions. As shown in Figure S8, the prompt structure is carefully designed to elicit temporally coherent reasoning that produces semantically valid negation captions grounded in the visual content.

Our prompt begins by informing the model that it will be provided with an image containing a highlighted bounding box, along with a target phrase describing the main subject in the region. The model is then guided through three distinct reasoning steps:

A.1.1 STEP 1: ATTRIBUTE EXTRACTION

The model first generates two comprehensive lists of attributes:

- **Present Attribute** (A_{pres}): At least three attributes or keyword items clearly visible within the bounded region.
- **Absent Attribute** (A_{abs}): At least three attributes or keyword items that are contextually relevant but clearly not present in the bounded region.

A.1.2 STEP 2: CAPTION GENERATION

Using the attributes from Step 1, the model produces two types of captions:

- **Negative Caption** (C_{neg}): Creates a factually incorrect statement by falsely claiming an existing attribute is absent. This caption must contain a negation expression (e.g., "no", "not", "without") coupled with an attribute from the existing contents list.
- **Positive Caption** (C_{pos}): Creates a factually correct statement by accurately describing an absent attribute as absent. This caption pairs a negation expression with an attribute from the absent contents list.

This approach yields contrastive pairs where the negative caption contradicts the visual evidence while the positive caption aligns with it, creating training data that specifically targets negation understanding.

A.1.3 STEP 3: SEMANTIC VERIFICATION

For quality assurance, each generated caption undergoes verification:

- **Negative Verification**: Confirms the caption (1) contains a negation expression, (2) references an existing attribute from Step 1, and (3) factually mismatches the actual content of the bounded region.

- **Positive Verification**: Confirms the caption (1) contains a negation expression, (2) references an absent attribute from Step 1, and (3) correctly describes the absence of the attribute in a way relevant to the context.

This verification step ensures semantic integrity and prevents generation artifacts by applying explicit logical checks. If either caption fails verification, the process iteratively regenerates captions until valid pairs are produced or the retry limit is reached.

The prompt enforces concise, natural language expressions with a single-sentence structure. As examples in Figure S9 and Figure S10, it requires the model to focus exclusively on the bounded region, preventing semantic drift to other parts of the image. The entire process outputs a structured JSON format containing the attribute lists, caption pairs, and verification rationales, facilitating downstream dataset creation and quality control processes.

A.2 VQA-BASED CAPTION ALIGNMENT

To address a critical challenge in negation-aware detection, ensuring generated captions reference exclusively the intended bounding box rather than other visually similar regions, we implement a structured verification pipeline with VQA alignment.

First, we apply alphabetical region labeling to all bounding boxes that share the target phrase type (e.g., "person") by assigning distinct markers (`A, B, C, ...`) to each instance. The originally prompted region remains unlabeled to avoid biasing the verification process. As shown in Figure S11, our visual prompting approach carefully considers label placement to maintain visual clarity. When labeling multiple instances of the same type (e.g., multiple "person" boxes), we position alphabetical markers outside the top-left corner of each bounding box to avoid occluding the object itself. This placement strategy preserves the visual integrity of the object while providing clear reference points for the VQA model. In cases where objects appear near image boundaries, we adaptively place labels inside the top-left corner of the bounding box to ensure they remain visible within the frame. This adaptive positioning is crucial for maintaining consistent label visibility across diverse image compositions.

Then, for each caption pair (C_{pos}, C_{neg}), we query a multimodal VQA model with two precisely formulated questions as in Figure S12. The VQA model analyzes the image and captions to produce structured JSON responses specifying matching box labels. A valid alignment requires that C_{pos} matches *exactly* the original unlabeled region, while C_{neg} either matches no regions (``None'') or incorrectly matches another box. This process effectively eliminates label noises: false negatives, where C_{neg} accidentally describes another instance, and ambiguous groundings, where captions generically describe multiple regions.

Figure S13 showcases several successful examples from our complete caption generation pipeline. In these examples, we can observe how the three-step CoT process first generates attribute-based negative and positive captions for the target region, followed by the VQA alignment step that verifies caption-region correspondence. Despite the effectiveness of our approach, we encountered certain limitations in complex scenes, as illustrated in Figure S14. When multiple instances of the same type are densely clustered, the visual prompting can become ambiguous, making it difficult for the VQA model to determine precise correspondences. To maintain dataset quality, we implemented a filtering mechanism that excludes images containing more than five instances of the same type from the caption generation process. This threshold was empirically determined to balance the diversity of the dataset with the precision of the annotation, ensuring that our training data provides unambiguous supervision signals for understanding the meaning of negations.

```
You are provided with an image in which the target object "<TARGET_PHRASE>" is highlighted using a red contoured
bounding box.  You are a vision-language model with advanced chain-of-thought reasoning.  You must produce both
negative and positive captions referencing the same main subject, "<TARGET_PHRASE>".

Step 1) Summarize the highlighted bbox existing/missing contents (color, action, location, relationship, shape,
texture, etc.):

    [Existing Contents] Provide at least 3 short attribute or keyword items that describe SHOWN within the red
    bounding box.
    - All contents should be CLEARLY CHECKED in image.
    - Example: If the region corresponds to 'woman', you could include items like ['running at left lane', 'brown
            hair', 'blue shirt', 'jumping', 'holding a bat'].
    [Absent Contents] Provide at least 3 short attribute or keyword items that describe NOT in the red bounding box.
    - All contents should be CLEARLY MISSING in image, but somewhat relvant to the situation.
    - Example: If the region corresponds to 'A woman in a blue shirt rides a bicycle', you could include items like
            ['helmet', 'glasses', 'red hoodie'], if all items are not in the image.

Step 2) For selected content items from step 1, produce exactly ONE negative caption and ONE positive caption with
negation expressions (e.g. 'no', 'not', 'never', 'without', 'un-', ...). Each caption should be about the bounding
box's main subject ("<TARGET_PHRASE>" in the red bbox) as the focus.

    [Negative caption]: Caption that mismatched with the target region by combining negation expression and existing
    content item.
    (1) Must contain a negation expression with Existing Contents.
    (2) Keep it a single sentence or phrase, but it can be descriptive on target region.
    (3) Example: If existing contents are ['man', 'blue shirt', 'hat'] -> select 'hat'
                => 'A man without hat on his head.' ('hat' with 'without')
                If existing contents are ['plate', 'on the top', 'black', 'near the woman']
                => select 'near the woman' => 'A black plate is not located near the woman.'

    [Positive caption]: Caption that match with target region containing absent concepts with negation expressions.
    (1) Must contain a negation expression with Absent Contents.
    (2) Keep it a single sentence or phrase, which is actually present or relevant.
    (3) Example: If absent contents are ['helmet', 'glasses', 'red hoodie'] => select 'red hoodie',
                you could say 'A woman without a red hoodie rides a bicycle.'

Step 3) Provide verification for each caption:

- After each negative or positive caption, include a short 'verification' string that clarifies why it is truly
  negative or positive, focusing on the use of the negation.
- Negative check: (1) Does it contain a negation expression? (2) Does it contain the existing item from Step 1? (3)
  Does it mismatch with the bounding box contents?
- Positive check: (1) Does it contain a negation expression? (2) Does it contain the absent item from Step 1? (3)
  Is that negation absent from the bounding box, but thematically relevant?


IMPORTANT:
- Keep each caption to one sentence. Natural, fluent English with a bit of descriptive detail is encouraged.
- Your bbox_contents and subsequent captions should provide unique or distinguishing details specifically about the
  object in the target region, ensuring that they do not unintentionally refer to objects or attributes that lie
  outside of this indicated region.
- Return your final answer in a JSON structure with the following schema:

{
        "steps": [ { "explanation": "...", "output": "..." }, ... ],
        "bbox_contents": {       "existing": [ ... ],      "absent":   [ ... ]  },
        "pairs": [
                {
                "content_item": {
                        "existing": "<one existing item>",
                        "absent":   "<one absent item>"
                },
                "negative_caption": "...",
                "negative_verification": "...",
                "positive_caption": "...",
                "positive_verification": "..."
                }
        ]
}

You should reveal your chain-of-thought in steps[1,2,3], but keep it concise and do not mention about visual prompt
in the final output sentences. Please identify at least 3 existing/missing items (other than the main subject) in
that region, then select one for generating negative/positive caption pairs with verification. Use the JSON schema
described above.
```

Figure S8: **Prompt for Three-step CoT Negation Caption Generation.** Our prompt guides the model to systematically (1) extract present and absent attributes from visually highlighted regions, (2) generate complementary negative and positive captions with explicit negation markers, and (3) verify semantic alignment through logical validation.

Input Image:

Target phrase:
"The boat"

STEP 1 — Summarize the highlighted bbox existing/missing contents

A_{pres}: ["dark silhouette", "still water reflection", "multiple antennas"]
A_{abs}: ["bright decorations", "passenger activity", "painted details"]

STEP 2 — For selected content items from step 1, produce negative and positive caption with negation expressions

C_{neg}: "The boat does **not** appear as a dark silhouette against the sunrise."
C_{pos}: "The boat is shown **without** any bright decorations along its structure."

STEP 3 — Provide verification for each caption

V_{neg}: "This caption uses 'not' with 'dark silhouette,' misrepresenting the boat's visual attribute."
V_{pos}: "This caption uses 'without' with 'bright decorations,' confirming its absence in the bounding box."

Input Image:

Target phrase:
"A man"

STEP 1 — Summarize the highlighted bbox existing/missing contents

A_{pres}: ["white beard", "blue cap", "standing"]
A_{abs}: ["gloves", "glasses", "apron"]

STEP 2 — For selected content items from step 1, produce negative and positive caption with negation expressions

C_{neg}: "A man **without** a blue cap on his head."
C_{pos}: "A man **without** glasses is standing in the doorway."

STEP 3 — Provide verification for each caption

V_{neg}: "It contains a negation ('without') and the existing item ('blue cap'), which mismatches because the man clearly wears a blue cap."
V_{pos}: "It contains a negation ('without') and the absent item ('glasses'), which is absent from the image and relevant to the setting."

Input Image:

Target phrase:
"The ball"

STEP 1 — Summarize the highlighted bbox existing/missing contents

A_{pres}: ["round", "black and white", "logo"]
A_{abs}: ["net", "grass", "foot touching"]

STEP 2 — For selected content items from step 1, produce negative and positive caption with negation expressions

C_{neg}: "The ball has **no** visible logo on it."
C_{pos}: "The ball shows **no** foot touching by any player on the field."

STEP 3 — Provide verification for each caption

V_{neg}: "Negative Verification: (1) Yes, it contains a negation. (2) Yes, it contains 'logo'. (3) Yes, it mismatches because the ball does have a logo."
V_{pos}: "Positive Verification: (1) Yes, it contains a negation. (2) Yes, it contains 'foot touching'. (3) Yes, it is absent but relevant to a ball."

Input Image:

Target phrase:
"a hat"

STEP 1 — Summarize the highlighted bbox existing/missing contents

A_{pres}: ["beige hat", "pattern", "wide brim"]
A_{abs}: ["letters", "passenger activity", "painted details"]

STEP 2 — For selected content items from step 1, produce negative and positive caption with negation expressions

C_{neg}: "The woman is **not** wearing a beige hat atop her head."
C_{pos}: "The hat on her head is **without** a letters."

STEP 3 — Provide verification for each caption

V_{neg}: "(1) Contains 'not' for negation. (2) Refers to 'beige hat'. (3) Contradicts the visual content, making it negative."
V_{pos}: "(1) Contains 'without' for negation. (2) Refers to 'letters' as absent. (3) The absence of a logo is true and contextually relevant."
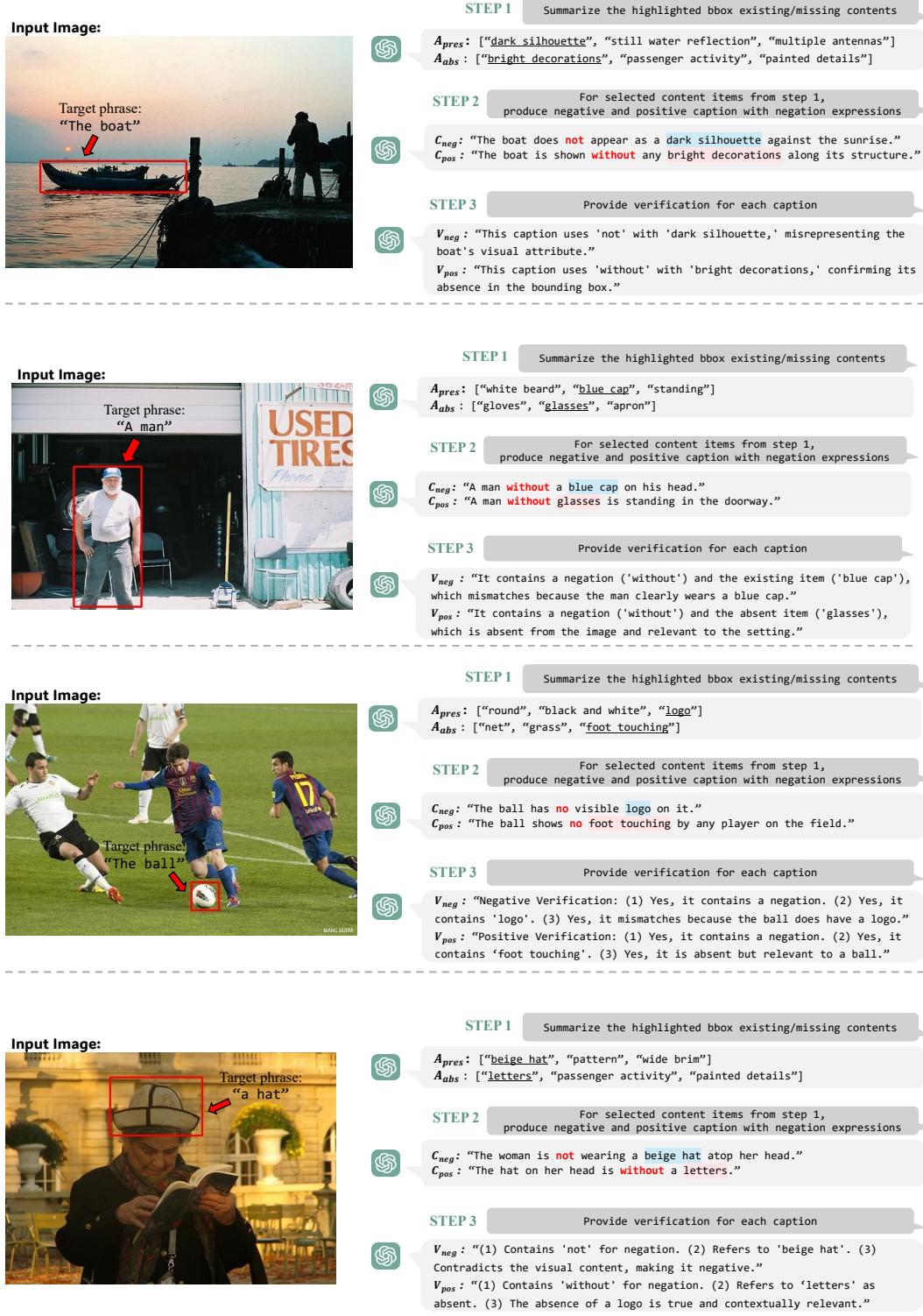
Figure S9: **Examples of CoVAND with 3-step CoT Caption Generation (1).** Example images and corresponding captions. Text with blue is present attribute(A_{pres}) and pink is absent attribute(A_{abs}). In detail, <negation word>+<A_{pres}> can generate negative caption(C_{neg}) and <negation word>+<A_{abs}> can generate positive caption(C_{pos}).

Input Image:

Target phrase: "A dog"

STEP 1 Summarize the highlighted bbox existing/missing contents

A_{pres}: ["wet fur", "wearing a collar", "looking upward"]
A_{abs}: ["clothing accessories", "a leash", "an obstacle"]

STEP 2 For selected content items from step 1, produce negative and positive caption with negation expressions

C_{neg}: "A dog that is **not** looking upward."
C_{pos}: "A dog **without** a leash is standing in the water."

STEP 3 Provide verification for each caption

V_{neg}: "This caption is negative as it uses a negation expression with an existing attribute, mismatching the observed orientation of the dog."
V_{pos}: "This caption is positive as it uses a negation expression for an absent attribute where no leash is visible, matching the lack of such an item."

Input Image:

Target phrase: "Infant Boy"

STEP 1 Summarize the highlighted bbox existing/missing contents

A_{pres}: ["blond hair", "blue clog sandals", "orange traffic cone"]
A_{abs}: ["hat", "glasses", "red hoodie"]

STEP 2 For selected content items from step 1, produce negative and positive caption with negation expressions

C_{neg}: "The infant boy is **not** wearing blue clog sandals."
C_{pos}: "The infant boy **without** a hat play with a traffic cone."

Provide verification for each caption

V_{neg}: "Contains negation 'not' and the existing item 'blue clog sandals', which is present, making this factually incorrect."
V_{pos}: "Contains negation 'without' and the absent item 'hat', which is indeed absent in the image."
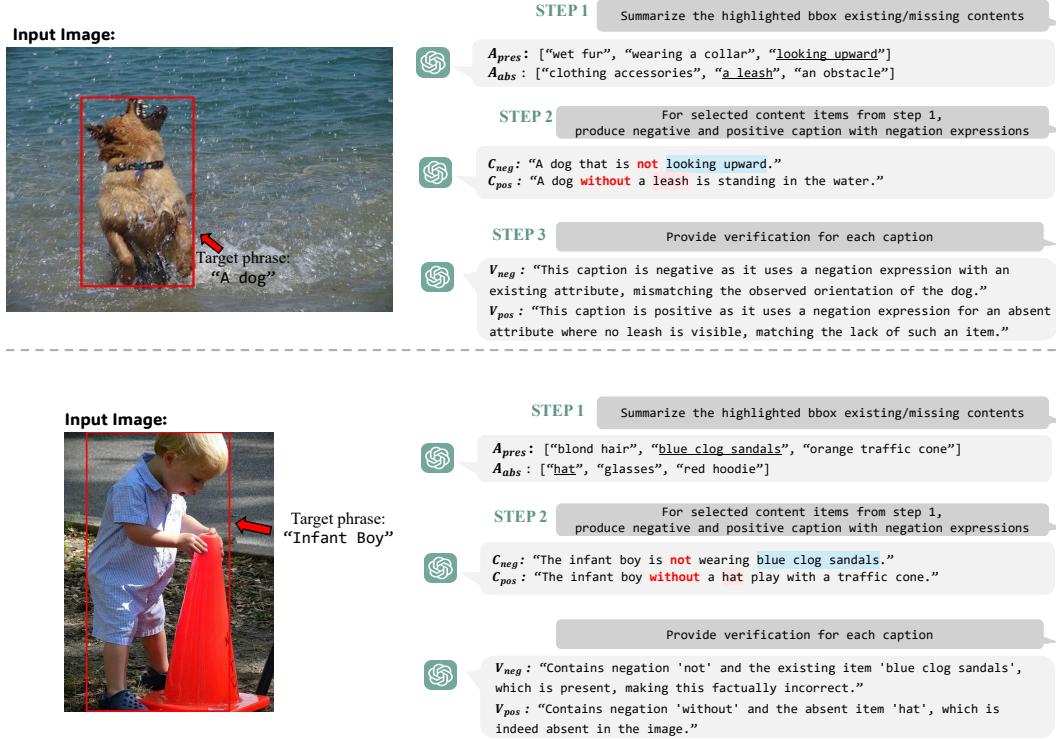
Figure S10: **Examples of CoVAND with 3-step CoT Caption Generation (2).** Example images and corresponding captions. Text with blue is present attribute(A_{pres}) and pink is absent attribute(A_{abs}). In detail, <negation word>+<A_{pres}> can generate negative caption(C_{neg}) and <negation word>+<A_{abs}> can generate positive caption(C_{pos}).

Figure S11: **Examples of Visual Prompt on VQA Alignments.** We apply alphabetical region labeling to all bounding boxes that share the target phrase type by assigning distinct markers (A, B, C, ...) to each instance with red bounding boxes.
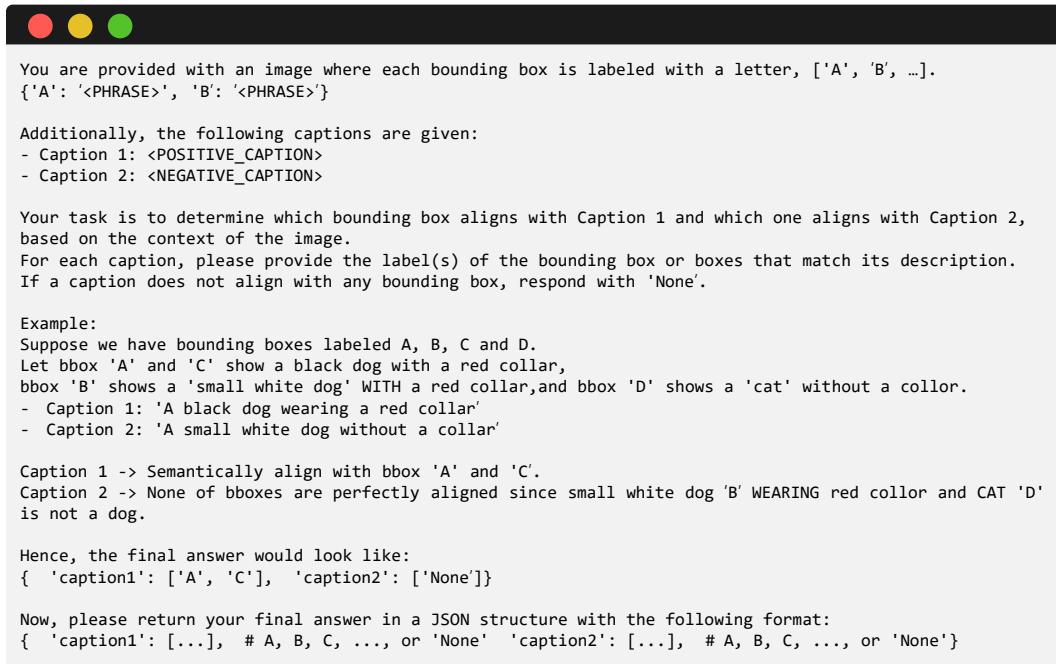
```
You are provided with an image where each bounding box is labeled with a letter, ['A', 'B', …].
{'A': '<PHRASE>', 'B': '<PHRASE>'}

Additionally, the following captions are given:
- Caption 1: <POSITIVE_CAPTION>
- Caption 2: <NEGATIVE_CAPTION>

Your task is to determine which bounding box aligns with Caption 1 and which one aligns with Caption 2,
based on the context of the image.
For each caption, please provide the label(s) of the bounding box or boxes that match its description.
If a caption does not align with any bounding box, respond with 'None'.

Example:
Suppose we have bounding boxes labeled A, B, C and D.
Let bbox 'A' and 'C' show a black dog with a red collar,
bbox 'B' shows a 'small white dog' WITH a red collar,and bbox 'D' shows a 'cat' without a collor.
- Caption 1: 'A black dog wearing a red collar'
- Caption 2: 'A small white dog without a collar'

Caption 1 -> Semantically align with bbox 'A' and 'C'.
Caption 2 -> None of bboxes are perfectly aligned since small white dog 'B' WEARING red collor and CAT 'D'
is not a dog.

Hence, the final answer would look like:
{ 'caption1': ['A', 'C'],  'caption2': ['None']}

Now, please return your final answer in a JSON structure with the following format:
{ 'caption1': [...],  # A, B, C, ..., or 'None'  'caption2': [...],  # A, B, C, ..., or 'None'}
```

Figure S12: **Prompt for VQA Alignment.** Our alignment process with (1) labeling all candidate bounding boxes with alphabetical markers, and (2) querying the VQA model to determine precise correspondences between generated captions and visually annotated regions.

Figure S13: **Examples of CoVAND.** Example images for the 3-step CoT Negation Caption Generation and the VQA alignment are needed. The VQA alignment step is only executed when there are multiple instances with the same phrase type.
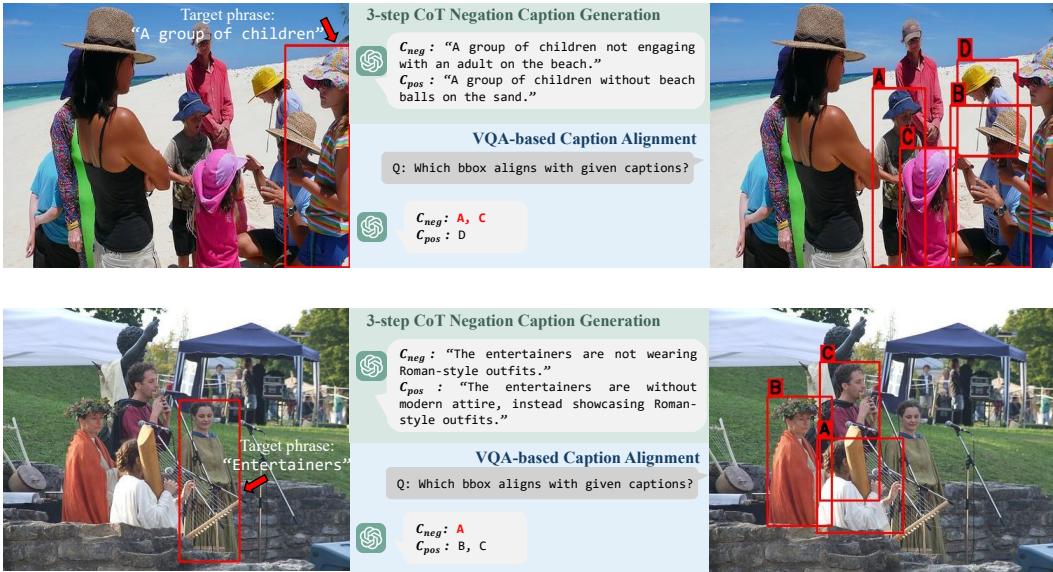
Figure S14: **Error on CoVAND.** VQA alignment occasionally fails when instances are densely clustered, making it difficult to determine which instance each visual prompt references.

B IMPLEMENTATION DETAILS

B.1 GROUNDING DINO MODEL

Our implementation is built upon the Grounding DINO architecture (Liu et al., 2024; Zuwei Long, 2023), which employs a dual-encoder-single-decoder design for vision-language understanding. For efficient fine-tuning towards negation understanding, we apply LoRA (Hu et al., 2021) to specific layers of the cross-modality decoder. The Grounding DINO consists of several key components:

- An image backbone (Swin Transformer (Liu et al., 2021)) for visual feature extraction
- A text backbone (BERT (Devlin et al., 2019)) for textual feature encoding
- A feature enhancer with self-attention and cross-attention mechanisms
- A language-guided query selection module that initializes query embeddings
- A cross-modality decoder that refines object detection based on both visual and text

We implement parameter-efficient fine-tuning by applying LoRA to `deep` layers (the final three cross-attention layers in the cross-modality decoder). This strategic placement allows us to modify how the model integrates negation cues from text with visual features while preserving pre-trained knowledge in earlier layers. Specifically, we insert LoRA only into the query (Q) and value (V) projections of the **text cross-attention**; the image deformable cross-attention and the self-attention blocks remain unchanged. The addition of ReLU activation between the down-projection and up-projection matrices, similar to (Chen et al., 2022), enhances the model's ability to capture non-linear relationships between negation cues and visual features. In Grounding DINO's cross-attention, the interactions operate as follows:

- **Image Cross-Attention**:
  - Query (Q): the updated cross-modality query from the preceding self-attention layer
  - Key (K) and Value (V): the image features processed through the feature enhancer
- **Text Cross-Attention**:
  - Query (Q): the output from the image cross-attention layer
  - Key (K) and Value (V): text features encoding language information

8

(a) Baseline Model

(b) Baseline Model + COVAND Fine-tuning (LoRA on `shallow` layers)

(c) Baseline Model + COVAND Fine-tuning (LoRA on `deep` layers)

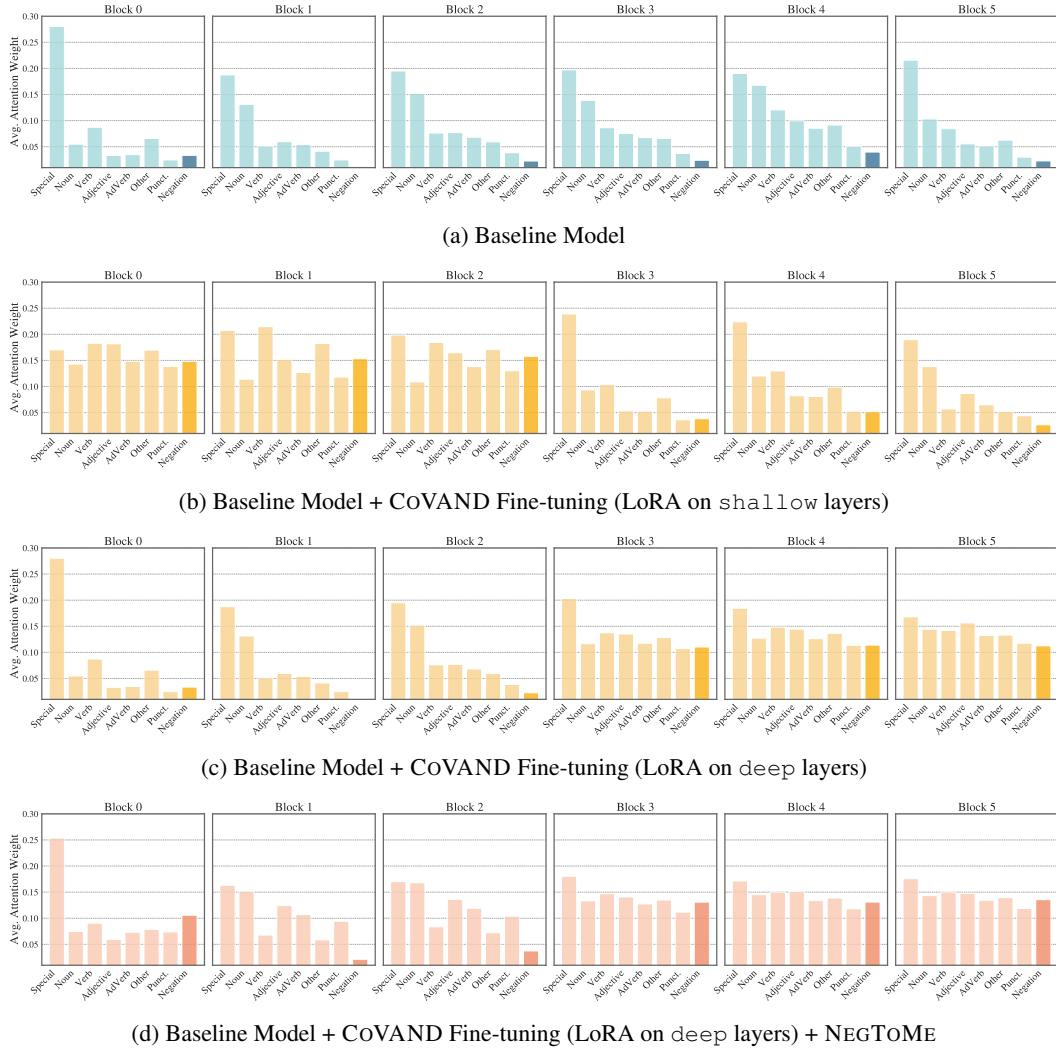(d) Baseline Model + COVAND Fine-tuning (LoRA on `deep` layers) + NEGTOME

Figure S15: **Average Attention Weights by Decoder Blocks.** We only update the LoRA modules while freezing other layers for fine-tuning. Placement of LoRA, `shallow` means LoRA located on early decoder blocks (0-2) and `deep` means LoRA located on latter decoder blocks (3-5).

Figure S15 reveals critical insights into the optimal placement of LoRA modules (Boenisch et al., 2025) for negation understanding. The baseline model (Figure S15a) shows a strong bias toward Special tokens across all decoder blocks, with negation cues receiving minimal attention. When we apply LoRA to `shallow` blocks (Figure S15b), negation tokens initially receive higher attention weights in blocks 0-2, but this effect rapidly diminishes in the later blocks where attention to negation drops.

In contrast, when we apply LoRA to `deep` blocks (Figure S15c), the model maintains consistent attention to negation tokens through blocks. This pattern persists through the final detection heads, explaining the superior negation-aware detection performance. Some works (Gao et al., 2025; 2024; Seputis et al., 2024) further validate our approach by demonstrating that allocation of adaptation capacity to mid-to-late transformer layers yields optimal results for complex semantic tasks.

With the addition of NEGTOME (Figure S15d), attention to negation tokens increases consistently across all blocks, with particular amplification in the final blocks where detection decisions are made. This confirms that our token merging strategy effectively preserves negation signals throughout the entire network, even in early blocks that did not receive LoRA adaptation. The combined effect creates a consistent processing path for negation cues from text encoding through to final

Table S5: **OVDEval-Negation Evaluation.** Performance on Grounding DINO tiny model.

	AP	NMS-AP (Yao et al., 2024)	FPR
G-DINO-T (Liu et al., 2024)	48.5	22.8	54.0
+ Ours	51.1	23.3	42.5
	(+2.6)	(+0.5)	(−11.5)

Table S6: **Trainable Parameter Ratio.** The table compares the total model size with the number of LoRA-tuned parameters for each detector and backbone pair. During fine-tuning on CoVAND, only the LoRA layers are trainable, with all other layers kept frozen with their pretrained weights.

	Image Backbone	Total Param.	LoRA Param.	Ratio (%)
G-DINO-T (Liu et al., 2024)	Swin-T (28.8M)	173M	8.2k	0.005
G-DINO-B (Liu et al., 2024)	Swin-B (88M)	233M	12.3k	0.005
APE-Ti (Shen et al., 2024)	ViT-Ti (5.8M)	771M	129k	0.017
APE-L (Shen et al., 2024)	ViT-L (307M)	1B	129k	0.012
Qwen-2.5-VL-3B (Bai et al., 2025)	ViT-H (632M)	3.8B	77k	0.002

detection, explaining the significant performance improvements observed in the OVDEval and D^3 benchmarks.

Together, these adaptations enable our model to effectively capture the semantics of negation by enhancing the cross-modal integration of negation cues with their corresponding visual attributes, resulting in more accurate detection under negation scenarios.

Compared with the tiny model of Grounding DINO baseline, we need merely 0.005% trainable parameters to capture negation cues effectively, as in Table S6. To keep the tiny model within the same 0.005% budget, we attach LoRA adapters to the 4 and 5 text cross-attention blocks of the decoder. Our lightweight adaptation yields a consistent performance gain: +2.6 AP and +0.5 NMS-AP, while slashing the False-Positive Rate (FPR) by **11.5%** as in Table S5. Although AP and NMS-AP improvements are moderate, they are achieved without sacrificing any metric; in fact, every reported score is on par with, or better than, the baseline, indicating that our negation-centric tuning does not degrade the detector's general ability.

B.2 APE MODEL

Our implementation builds upon the APE framework (Shen et al., 2024), a universal visual perception model that unifies detection, segmentation, and grounding through instance-level region-sentence alignment. The architecture features several key innovations:

- A vision backbone (ViT-L (Dosovitskiy et al., 2020)) pretrained with EVA-CLIP (Sun et al., 2023) for visual feature extraction
- A text encoder (EVA02-CLIP (Fang et al., 2024)) processing both categorical vocabularies and free-form descriptions
- A gated cross-modality interaction module that fuses visual and text features
- A transformer decoder with deformable attention (Zhu et al., 2020) for joint reasoning

APE introduces a novel gated fusion mechanism that efficiently handles thousands of prompts per forward pass. Unlike previous approaches that directly fuse all text features (Li et al., 2022), APE implements conditional interaction paths:

$$\hat{V} = \begin{cases} V + \text{Attn}(V, P_{\text{voc}}) & \text{for vocabulary prompts} \\ \text{Attn}(V, P_{\text{sen}}) & \text{for sentence descriptions} \end{cases} \tag{4}$$

where V denotes visual features and P represents text embeddings. This gating strategy reduces FLOPs compared to GLIP-style fusion (Li et al., 2022). The model processes inputs at 1,024 pixel

resolution using AdamW optimization Loshchilov & Hutter (2017) with learning rate 0.0005 and weight decay 0.05. We employ large-scale jittering augmentation (Ghiasi et al., 2021) with random scales from 0.1 to 2.0. We train APE-Ti models with four A6000 GPUs with a batch size of 4.

We apply LoRA exclusively to the encoder's cross-attention layers where visual and text features interact. This targeted adaptation modifies only 0.017% of APE's parameters as in Table S6. Despite APE-L's strong theoretical performance, its 1B parameters exceed the 48GB memory capacity of NVIDIA A6000 GPUs during training. We therefore focus on APE-Ti, which achieves 32.5 AP on D^3 while maintaining practical deployability.

B.3 QWEN-2.5-VL MODEL

In addition to dedicated detectors, we test our method's generalizability on a powerful Multimodal Large Language Model (MLLM), Qwen-2.5-VL (Bai et al., 2025). Unlike dual-encoder architectures, Qwen-2.5-VL is an end-to-end model that directly processes interleaved image and text data. As detailed in its technical report, the architecture consists of three main components:

- A Vision Transformer (ViT-H) that is redesigned and trained from scratch to handle native resolution inputs. For efficiency, it incorporates windowed attention in most layers, with full self-attention only in specific blocks. The ViT architecture is also updated with RMSNorm and SwiGLU activations to align with modern LLM design principles.
- An MLP-based Vision-Language Merger that compresses spatially adjacent patch features before feeding them into the language model, enhancing computational efficiency.
- A Large Language Model decoder based on the Qwen2.5 architecture, which performs unified reasoning over the combined multimodal input and generates responses, including object coordinates for detection tasks.

For parameter-efficient fine-tuning, we again employ LoRA, strategically targeting the `deep` layers of the LLM decoder to enhance its negation reasoning without disturbing its foundational knowledge. Based on our experimental setup, LoRA adapters are specifically injected into the query (`q_proj`) and value (`v_proj`) projections of the self-attention modules within layers (15, 24, 30). This targeted placement is designed to modulate how the model integrates visual information with textual negation cues in its higher-level semantic reasoning stages. We configure the LoRA adapters with a rank r of 4 and a dropout probability of 0.05.

Following the execution script, the Qwen-2.5-VL-3B model is fine-tuned for 1 epoch on our COVAND dataset on H200 GPU. We use a learning rate of $5e - 5$ with the AdamW optimizer (Loshchilov & Hutter, 2017) and a per-device batch size of 32, with 2 gradient accumulation steps, totaling an effective batch size of 64. The model is trained using bfloat16 mixed-precision. During this process, all original model parameters-including the ViT, MLP merger, and LLM backbone—are kept frozen; only the injected LoRA adapter weights are updated.

C COMPARISON WITH POST-HOC VQA METHODS

Motivation. An alternative to enhancing a detector's internal negation understanding is a two-stage pipeline, where a standard detector generates initial proposals and a powerful Multimodal Large Language Model (MLLM) then acts as a post-hoc filter to remove erroneous detections. To investigate the viability and trade-offs of this common alternative, we implemented two post-hoc VQA variants. We built these on top of the same baseline detector used in our main experiments and report results on the OVDEval Negation subset using both AP and class-ignored NMS-AP.

Two post-hoc settings. *(A) Crop & Verify.* For each image, we take the detector's top-k boxes, crop each region, and query an MLLM with a yes/no question about whether the crop satisfies the input description. This yields k separate MLLM calls per image. *(B) Coordinate Prompting.* We avoid cropping and instead pass all top-k box coordinates and the description to the MLLM at once, asking it to indicate which boxes are inconsistent.

Results. As shown in Table S7, the *Crop & Verify* method substantially increases the baseline detector's NMS-AP from 36.8 to 54.4, confirming that a strong VQA filter can reduce contradictory

Table S7: **Post-hoc VQA on OVDEval Negation.** Numbers are AP / NMS-AP (↑). "Ours" is the single-stage detector fine-tuned with deep-layer LoRA + NEGToME. Crop & Verify improves the baseline but requires k MLLM calls per image; Coordinate Prompting is faster but brittle. Stacking the expensive verifier on top of *our* improved detector yields the best overall numbers.

Detector	Post-hoc Verifier	AP	NMS-AP
(A) Crop & Verify (top-k crops $\Rightarrow k$ MLLM calls)			
G-DINO-B		54.0	36.8
G-DINO-B	+ Qwen-2.5-VL-3B	59.2	54.4
G-DINO-B + **Ours**		58.7	44.5
G-DINO-B + **Ours**	+ Qwen-2.5-VL-3B	**63.8**	**58.4**
(B) Coordinate Prompting (single MLLM call with all boxes)			
G-DINO-B		54.0	36.8
G-DINO-B	+ Qwen-2.5-VL-3B	48.6	34.1
G-DINO-B	+ Qwen-2.5-VL-7B	54.0	36.9
G-DINO-B + **Ours**		58.7	**44.5**
G-DINO-B + **Ours**	+ Qwen-2.5-VL-3B	49.6	37.0
G-DINO-B + **Ours**	+ Qwen-2.5-VL-7B	**58.6**	44.4

detections. However, this accuracy gain comes with a heavy latency cost due to the $O(k)$ MLLM calls required per image. In contrast, the faster *Coordinate Prompting* method is unreliable for fine-grained reasoning and often degrades performance; for instance, the baseline's NMS-AP drops from 36.8 to 34.1 when paired with the 3B verifier. Notably, our single-stage method already achieves an NMS-AP of 44.5, closing much of this performance gap without any added latency. When the accurate but slow *Crop & Verify* filter is applied on top of our already-improved model, it achieves the highest NMS-AP of 58.4, indicating that our method and post-hoc verification are complementary rather than redundant.

Conclusion. These experiments demonstrate that while a two-stage VQA pipeline can be effective, it presents a clear trade-off between accuracy and speed. The crop-based verifier is accurate but slow, whereas coordinate prompting is fast but brittle. Our single-stage approach, by contrast, instills negation sensitivity directly within the detector, improving the stricter NMS-AP and reducing false positives in a single, efficient pass. This confirms that post-hoc filtering does not obviate the need for a negation-aware detector. For practical, real-time settings, integrating negation reasoning directly into the model's fusion layers remains the most effective path. If latency is not a concern, our work also shows that a costly verifier can be used to further refine the outputs of our model.

D EVALUATION ON FULL OVDEVAL SUBSETS

OVDEval (Yao et al., 2024) is a comprehensive benchmark designed to evaluate the generalization capability of open-vocabulary detection (OVD) models across diverse linguistic aspects. The dataset includes 9 sub-datasets that test 6 distinct aspects: object, proper noun (landmark, logo, celebrity), attribute (color, material), position, relationship, and negation. Each subset features meticulously curated hard negative samples that challenge models to demonstrate true understanding of fine-grained linguistic descriptions rather than exploiting dataset biases. For instance, the color subset includes negative labels with the same object category but different colors, while relationship subsets maintain identical subjects and objects but alter the connecting verbs.

D.1 THE INFLATED AP PROBLEM AND NMS-AP METRIC

Standard Average Precision (AP) metrics face limitations when evaluating fine-grained described object detection due to what OVDEval terms the *Inflated AP Problem*. This issue occurs when a model predicts multiple bounding boxes for the same object with different labels, including mutually exclusive ones as in Figure S16. For example, a model might predict both "outdoor dog led by rope" and "dog not led by ropes outside" for the same dog, artificially inflating its AP score. Mathematically, this manifests as:
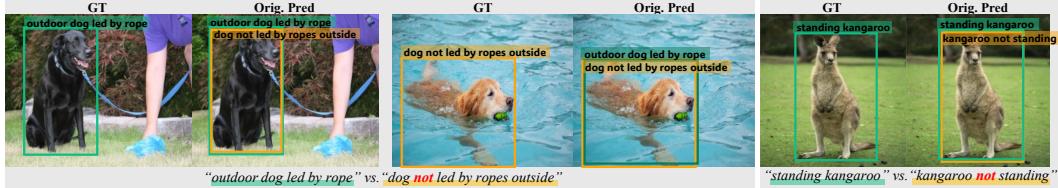
Figure S16: Failure Cases of Prior Models on Negation Descriptions

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{1}{1 + 1} = 0.50, \quad \text{Recall} = \frac{TP}{GT_{num}} = \frac{1}{1} = 1.0 \qquad (5)$$

Where a model with no actual understanding of attributes can still achieve a mAP of 0.50. To address this, we follow OVDEval's Non-Maximum Suppression Average Precision (NMS-AP) metric (Yao et al., 2024), which applies class-ignored NMS to remove redundant predictions for the same object before AP calculation. This provides a more accurate assessment of a model's ability to understand fine-grained descriptions of contradictory pairs.

D.2 GENERALIZATION TO NON-NEGATION SUBSETS

Table S8 demonstrates that our model maintains robust performance across all OVDEval subsets despite being trained exclusively on the negation-focused CoVAND dataset. Notably, our approach shows improved NMS-AP scores for Logo (+0.2), Landmark (+4.8), Color (+0.7), and Relationship (+3.8) subsets compared to the baseline. This broad generalization suggests that our negation-sensitive adaptations enhance the model's overall reasoning capabilities for complex descriptions. These results confirm that our LoRA-based parameter-efficient fine-tuning and NEGTOME token merging strategy provide benefits beyond negation understanding, enhancing the model's capability to process compositional descriptions across multiple semantic aspects.

Table S8: **Evaluation Results on Full OVDEval.** Performance on OVDEval subsets, except for the Negation. Even though we only trained with negation-focused CoVAND dataset, our models show robust results for other subsets.

	Logo		Landmark		Celebrity		Color		Material		Position		Relationship		Average	
	AP	NMS-AP	AP	NMS-AP	AP	NMS-AP	AP	NMS-AP	AP	NMS-AP	AP	NMS-AP	AP	NMS-AP	AP	NMS-AP
G-DINO	11.7	7.6	20.5	16.5	6.7	0.8	7.9	5.6	15.2	5.5	74.7	60.6	41.3	18.3	25.4	16.4
+Ours	11.5	7.8	22.4	21.3	6.6	0.3	7.9	6.3	15.8	5.3	70.5	54.6	42.3	22.1	25.2	16.8

E   ANALYSIS ON RPN-BASED DETECTOR

E.1   LIMITATIONS OF RPN-BASED DETECTORS UNDER NEGATION

Marginal or negative gains with LoRA.   Two–stage region–proposal detectors such as GLIP (Li et al., 2022) and FIBER (Dou et al., 2022) obtain slight improvement on negation–focused benchmarks after attaching LoRA adapters as in Table S9. For GLIP, whose backbone consists of stacked multi–head attention blocks, we inject LoRA only into the FFN layers of the last two transformer blocks, leaving all attention projections frozen. Even with this targeted fine-tuning, the gains remain marginal. These findings indicate that low-rank fine-tuning brings far smaller gains to GLIP than to DETR-style detectors. The gap can be traced to their attention layouts: GLIP employs self-MHA over a mixed token pool, whereas Grounding-DINO uses two modality-specific cross-attention blocks driven by a compact query set, a design that lets LoRA and NEGTOME act directly on phrase-level cues and thus respond much more strongly to negation.

Affirmative bias and context insensitivity.   Recent negation benchmarks reveal a sharp drop in detection accuracy whenever a query expresses absence or negation (Alhamoud et al., 2025). GLIP

and FIBER often treat a negated phrase (*"not X"*) as if it were *"X"*, triggering on object names while ignoring context qualifiers. Consequently, GLIP still localizes a *"microphone"* when the description states *"a person with no microphone"*, producing hallucinated objects. LoRA-adapted RPN detectors exhibit diminishing returns on negation-centric tasks because their proposal stage detects any region matching a noun, leaving little capacity to encode absence semantics.

Performance gap between AP and NMS-AP on the Negation subset. Table S9 further shows that model capacity alone does not resolve the issue: even the larger GLIP-L still exhibits a gap between AP and class-ignored NMS-AP, substantially wider than the gap of smaller DETR counterparts. The gap quantifies how many redundant, mutually exclusive boxes each model produces. A large drop after class-ignored NMS indicates that the detector continues to fire on the noun even when the query contains a negation cue, confirming the affirmative bias analyzed in the main paper.

Effect of NEGTOME. DETR keeps token granularity throughout the vision–language stack, allowing a merged phrase embedding to dominate $\langle q, k_i \rangle$ for its specific key k_i while leaving other keys unaltered. By contrast, GLIP or FIBER fuse language either by (a) global pooling of the entire caption (`[cls]`), or (b) class-name pooling plus a separate visual prompt. Both strategies erase intra-sentence polarity (*dog* vs. *not dog*) before the detector sees it. Token merging cannot recover that lost contrast; at best it shortens a sequence that will be pooled anyway.

Table S9: **OVDEval-Negation Evaluation on Additional Architectures.** RPN-based detectors show a large gap between AP and NMS-AP. FT denotes fine-tuning of LoRA parameters only, with all other pretrained weights kept frozen as in the main paper. Results marked with [†] are reproduced.

	Image Backbone	Total Param.	AP	NMS-AP (Yao et al., 2024)
Non RPN-based Detector				
MDETR (Kamath et al., 2021)	ResNet-101	185M	41.1	28.3
OmDet (Zhao et al., 2024b)	ConvNext-B	242M	55.9	35.1
Grounding DINO[†] (Liu et al., 2024)	Swin-B	233M	54.0	36.8
RPN-based Detector				
FIBER (Dou et al., 2022)	Swin-B	252M	57.2	28.7
GLIP-L (Li et al., 2022)	Swin-L	430M	51.8	29.3
GLIP-T (Li et al., 2022)			47.7	25.4
+ FT w.CoVAND	Swin-T	232M	47.8	26.1
+ FT w. CoVAND + NEGTOME			48.3	26.0

E.2 ADVANTAGES OF DETR–STYLE DETECTORS WITH LORA

Compositional reasoning. DETR-style detectors with transformer decoders (Kamath et al., 2021; Liu et al., 2024; Li et al., 2023; Shen et al., 2024) perform joint text–image reasoning through cross-attention in decoder blocks. This design enables natural handling of relations such as *"X but not Y"*.

Effectiveness of LoRA. Injecting LoRA adapters into the decoder cross-attention layers of Grounding DINO and fine-tuning on a negation-focused dataset improves mAP by $+2.6$ and cuts the false positive rate by 11.5%. The same lightweight adaptation reduces spurious detections on the Negation subset of OVDEval by nearly half, while preserving general detection accuracy.

Why the architecture helps. Each decoder layer attends to textual tokens; negation words therefore, modulate visual attention directly. In RPN pipelines, language supervision is applied only after proposals are fixed, limiting early rejection of forbidden objects. A fully fused DETR decoder yields a contextual representation of "what *not* to detect," which a small LoRA module can efficiently refine.

Advantage of NEGTOME. Both Grounding DINO and APE inherit the sub-token fragmentation of their text backbones with BERT and BPE in CLIP. NEGTOME merges those fragments into one polarity–aware phrase embedding and re-weights it by a boost factor β. In Grounding DINO, this merged vector is fed intact through token-level cross-attention, so every decoder layer receives a sharper gradient signal for the absence condition; the result is a $+10.8$ rise in NMS-AP and a 19.1% drop in false positives on OVDEval-Negation. APE employs CLIP, whose text encoder pools all

tokens into a single sentence vector before fusion. Here NEGTOME acts pre-pooling: by assigning larger softmax weights to the merged negation phrase it skews the sentence representation toward the correct polarity, yet does not increase sequence length. Consequently, the lightweight merger lifts APE-Ti by $+1.2$ in NMS-AP and reduces absent-object errors by 8.3%, despite updating only 0.017% of parameters. NEGTOME aligns with the inductive bias of both encoders: it supplies BERT-based decoders with an explicit token for cross-modal attention, and it biases CLIP's global pooling toward the correct semantic polarity. The mechanism is encoder-agnostic and therefore complements LoRA across heterogeneous DETR frameworks.

DETR-based detectors fine-tuned with LoRA and NEGTOME achieve larger and more reliable gains on negation and other compositional queries than RPN counterparts. Their set-prediction decoder offers a single, expressive locus for parameter-efficient language adaptation.

F ZERO-SHOT DOWNSTREAM TASKS: MULTIPLE CHOICE QUESTIONS

To further analyze our model's semantic comprehension of negation, we evaluate it on the NegBench Multiple Choice Question (MCQ) benchmark (Alhamoud et al., 2025). This benchmark is specifically designed to diagnose a VLM's ability to handle negation by requiring it to select the most accurate caption for an image from four options. These options are structured into three challenging categories as detailed below, providing a fine-grained analysis of a model's capabilities.

F.1 STRUCTURE OF THE NEGBENCH MCQ

The NegBench MCQ task (Alhamoud et al., 2025) generates multiple-choice questions where one answer is correct and the other three serve as hard negatives, designed to mislead models that do not properly understand negation. The questions are categorized into three distinct types based on the linguistic structure of the correct answer:

- **Positive Subset:** The correct caption is a simple affirmation that accurately describes objects present in the image (e.g., "*This image shows a baseball bat and baseball glove*"). This subset tests the model's fundamental visual grounding capabilities, as shown in Figure S17. Incorrect options often involve falsely negating a present object.

- **Negative Subset:** The correct caption accurately negates the presence of an object that is contextually relevant but absent from the image (e.g., "*A bowl is not present in this image*"). This directly tests the model's ability to comprehend explicit negation, as illustrated in Figure S18.

- **Hybrid Subset:** The correct caption combines both an affirmation and a negation within a single sentence (e.g., "*This image features a person, with no truck in sight*"). As shown in Figure S19, this is the most challenging subset as it requires compositional reasoning and an understanding of complex sentence structures that assign different polarities to different objects.

F.2 ERROR PATTERN ANALYSIS OF BASELINE MODELS

Our qualitative analysis reveals that baseline models exhibit consistent and fundamental error patterns on the NegBench MCQ task, primarily stemming from a severe *affirmative bias* as below:

1. **Blatant Contradiction of Visual Facts:** The most common failure is choosing a caption that directly contradicts the visual evidence. For example, in Figure S18, the baseline model selects "*There is no horse in this image*" for an image clearly depicting a horse. This indicates that the model heavily weighs the noun ("*horse*") while effectively ignoring the negation cue, treating both affirmative and negative statements as semantically similar.

2. **Polarity Confusion in Hybrid Sentences:** In the Hybrid subset (Figure S19), baseline models systematically fail to parse sentences containing both positive and negative clauses. For instance, given the ground truth "*This image features a refrigerator, but lack of a bottle*," the baseline chooses "*This image features a bottle, but does not include a refrigerator*." This shows a critical failure in compositional reasoning, where the model cannot correctly assign presence and absence to different objects within the same logical construct.

3. **Selection of Suboptimal Negatives:** In some cases on the Negative subset, the baseline avoids direct contradiction but fails to select the most accurate description. As seen in Figure S18, when the ground truth is "*A bowl is not present*," the baseline chooses "*no cake is present*." While factually correct, this choice suggests the model lacks a deeper contextual understanding to identify the most salient absent object among multiple true negative options.

These error patterns underscore that many state-of-the-art VLMs do not understand negation. Instead, they rely on shortcut strategies that collapse the semantic meaning of affirmative and negative statements. This motivates the need for methods that can fundamentally address this architectural limitation.

Figure S17: **Qualitative Results on the Positive subset of the Multiple Choice Question benchmark.** Captions with green checkmark ✅ is **GT**, pink refer to **Baseline**, and blue refer to **Ours**.

Figure S18: **Qualitative Results on the Negative subset of the Multiple Choice Question benchmark.** Captions with green checkmark ✅ is **GT**, pink refer to **Baseline**, and blue refer to **Ours**.

(1) This image features a refrigerator, but lack of a bottle.

(2) This image features a bottle, but does not include a refrigerator.

(3) A bottle is included in this image.

(4) A refrigerator is not present in this image.

(1) This image includes a sheep but no truck.

(2) This image shows a truck, but there is no sheep.

(3) This image shows a truck.

(4) There is no sheep in this image.

(1) This image features a dining table, but no car is present.

(2) This image features a car, but does not include a dining table.

(3) A car is included in this image.

(4) A dining table is not included in this image.

(1) This image features a frisbee, but there's no bench in sight.

(2) This image features a bench, but no frisbee is visible.

(3) This image shows a bench.

(4) A frisbee is not included in this image.

Figure S19: **Qualitative Results on the Hybrid subset of the Multiple Choice Question benchmark.** Captions with green checkmark ✅ is **GT**, pink refer to **Baseline**, and blue refer to **Ours**.

G   QUALITATIVE RESULTS

We present additional qualitative examples from the OVDEval and D^3 datasets to further demonstrate the effectiveness of our negation understanding approach. Figure S20 and Figure S21 show our model's ability to distinguish between contradictory attribute pairs such as "horse urinating" versus "horse that is not urinating" and "complete pizza" versus "pizza that is not complete". The baseline model often detects identical regions for both negative and positive descriptions, demonstrating significant affirmative bias. In contrast, our method successfully differentiates between these contradictory descriptions by correctly emphasizing negation cues.

Figures S22 to S24 illustrate our model's performance on the D^3 dataset. For descriptions such as "hanger without clothes" and "a bed without patterns", our model correctly identifies only the objects that satisfy these negated constraints. The baseline frequently exhibits false positives by detecting objects regardless of negation markers. Our approach demonstrates particular effectiveness for simple negation cases involving physical attributes and object presence.

Despite these improvements, our method still exhibits limitations in scenarios requiring highly complex reasoning, as shown in Figure S25. These challenges often involve multi-step relational logic combined with negation, such as in the query "a woman in white wedding dress not beside any men in suits", or understanding negated states, as in "a volley ball in the middle of the air untouched". Furthermore, resolving ambiguous or implicit negation cues like "unlike" in "origami unlike bird" remains a difficult problem. A common failure pattern in these cases is that when a complex event or state is entirely absent from the image (e.g., "the person who was proposed to on one knee"), the model defaults to its affirmative bias, detecting the main subject of the query ("person") rather than correctly identifying that no object matches the full description. Crucially, the baseline model faces identical challenges in these cases, demonstrating that these are open problems for the current generation of VLM detectors. This confirms that our method, while not a complete solution for such intricate reasoning, does not degrade performance on these hardest examples. These limitations highlight important areas for future research in handling complex linguistic constructions and multi-step negation scope resolution.

H   DECLARATIONS

LLM usage.   A large language model (LLM) was used during the preparation of this paper to proofread and refine the writing, including correcting grammar and improving sentence structure.

Ethics Statement.   Our work adheres to the ICLR Code of Ethics. The primary goal of this research is to improve the reliability and safety of vision-language models by addressing a fundamental flaw in their reasoning—the failure to understand negation. By reducing "affirmative bias," we aim to create models that align more closely with human language and intent, which can prevent critical errors in real-world applications (e.g., medical imaging or autonomous systems). Our new dataset, COVAND, is built upon the public Flickr30k Entities benchmark. The new captions are generated using a large language model (GPT-4o) with a systematic, multi-step pipeline designed to ensure high-quality, relevant, and grounded annotations. While our method improves a model's linguistic comprehension, it does not inherently address or remove societal biases that may be present in the underlying web-scale pre-training data or the baseline models themselves. We believe the contribution is a net positive, leading to more robust and predictable AI systems.

Reproducibility Statement.   We are committed to ensuring the reproducibility of our research. To this end, we will make our source code, including the implementation of the NEGTOME module, and the complete COVAND dataset publicly available upon publication.

Figure S20: **Qualitative Results on OVDEval Datasets (1).** Prediction results on contradictory caption pairs (yellow box vs. green box) from the negation subset of OVDEval dataset. Each row displays (left) ground-truth boxes, (middle) baseline predictions, and (right) our predictions. Our model effectively reduces affirmative bias, no longer returning identical bounding boxes for captions that express opposite meanings.
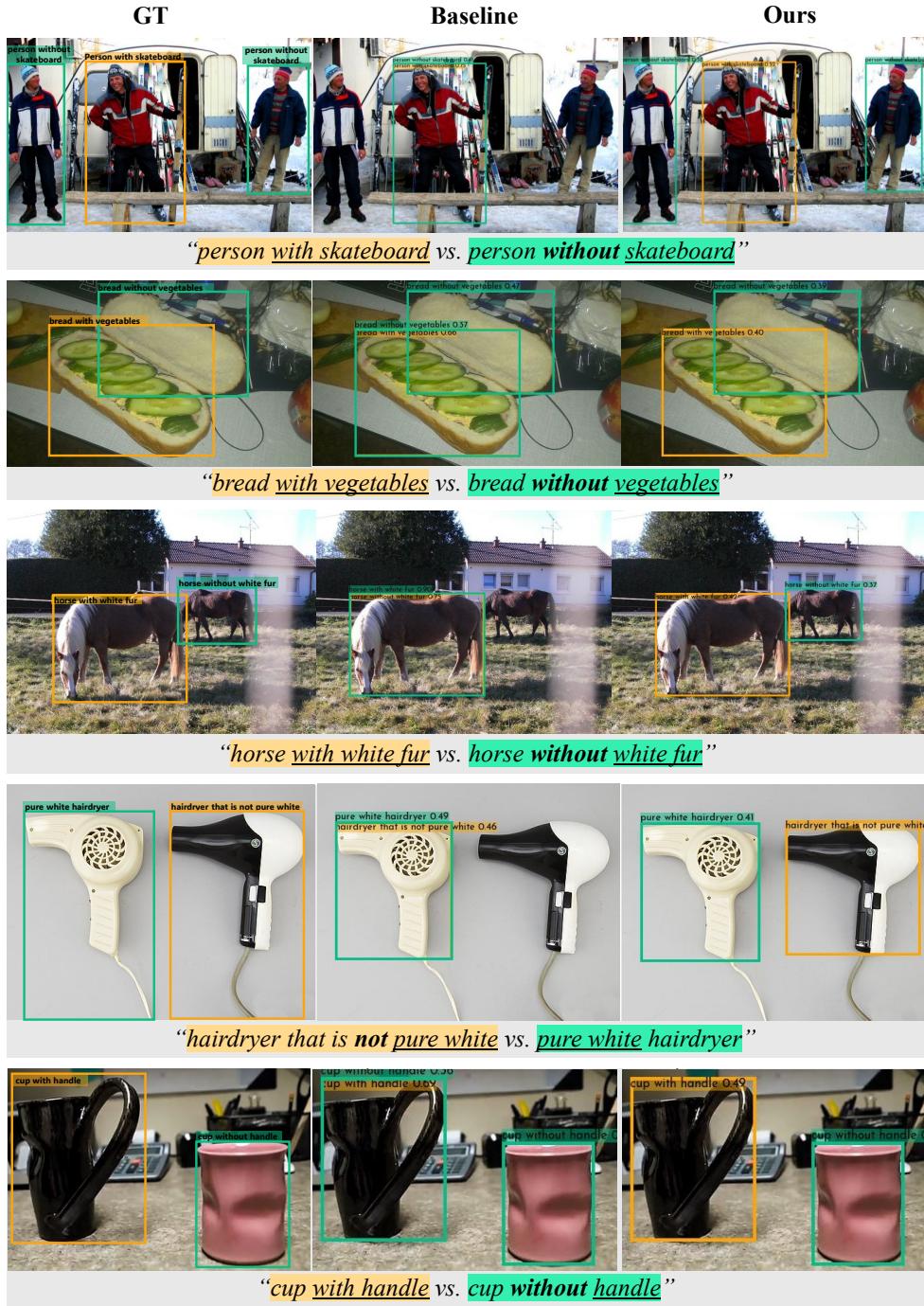
GT   Baseline   Ours

*"person with skateboard vs. person **without** skateboard"*

*"bread with vegetables vs. bread **without** vegetables"*

*"horse with white fur vs. horse **without** white fur"*

*"hairdryer that is **not** pure white vs. pure white hairdryer"*

*"cup with handle vs. cup **without** handle"*

Figure S21: **Qualitative Results on OVDEval Datasets (2).** Prediction results on contradictory caption pairs (yellow box vs. green box) from the negation subset of OVDEval dataset. Each row displays (left) ground-truth boxes, (middle) baseline predictions, and (right) our predictions. Our model effectively reduces affirmative bias, no longer returning identical bounding boxes for captions that express opposite meanings.

GT	Baseline	Ours

"clothed dog"

*"hanger **without** clothes"*

*"a bed **without** patterns on it"*

*"horseman **without** helmet"*

*a person at the conference table who is **not** seated*
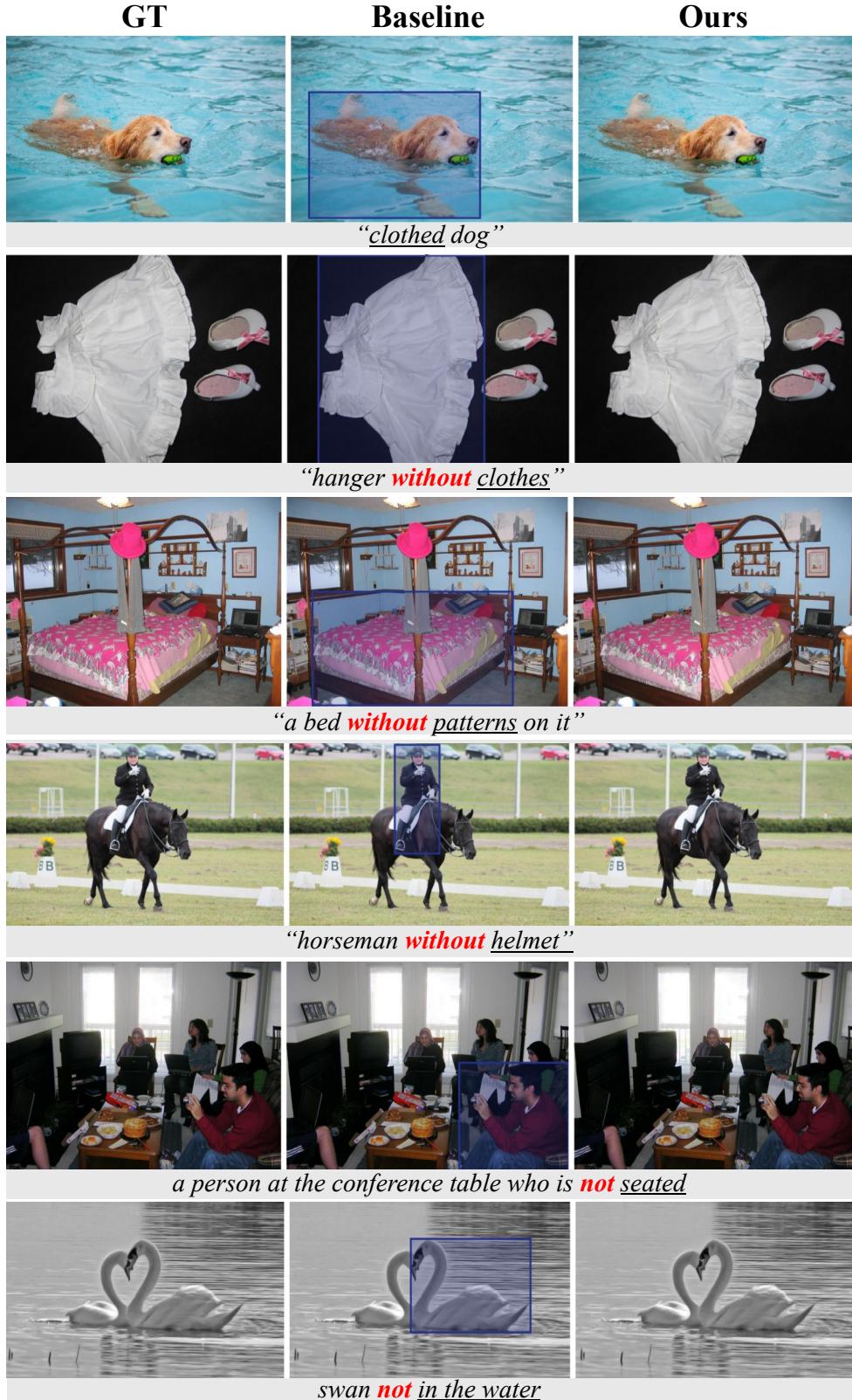
*swan **not** in the water*

Figure S22: **Qualitative Results on D³ Datasets (1).** Absence of a bounding box shows the model has determined that no instance in the image matches the input description. By filtering out such invalid predictions, our approach reduces affirmative bias and lowers the false-positive rate.

Figure S23: **Qualitative Results on D³ Datasets (2).** Absence of a bounding box means the model has determined that no instance in the image matches the input description. Our model effectively reduces the affirmative bias while keeping the correct predictions.
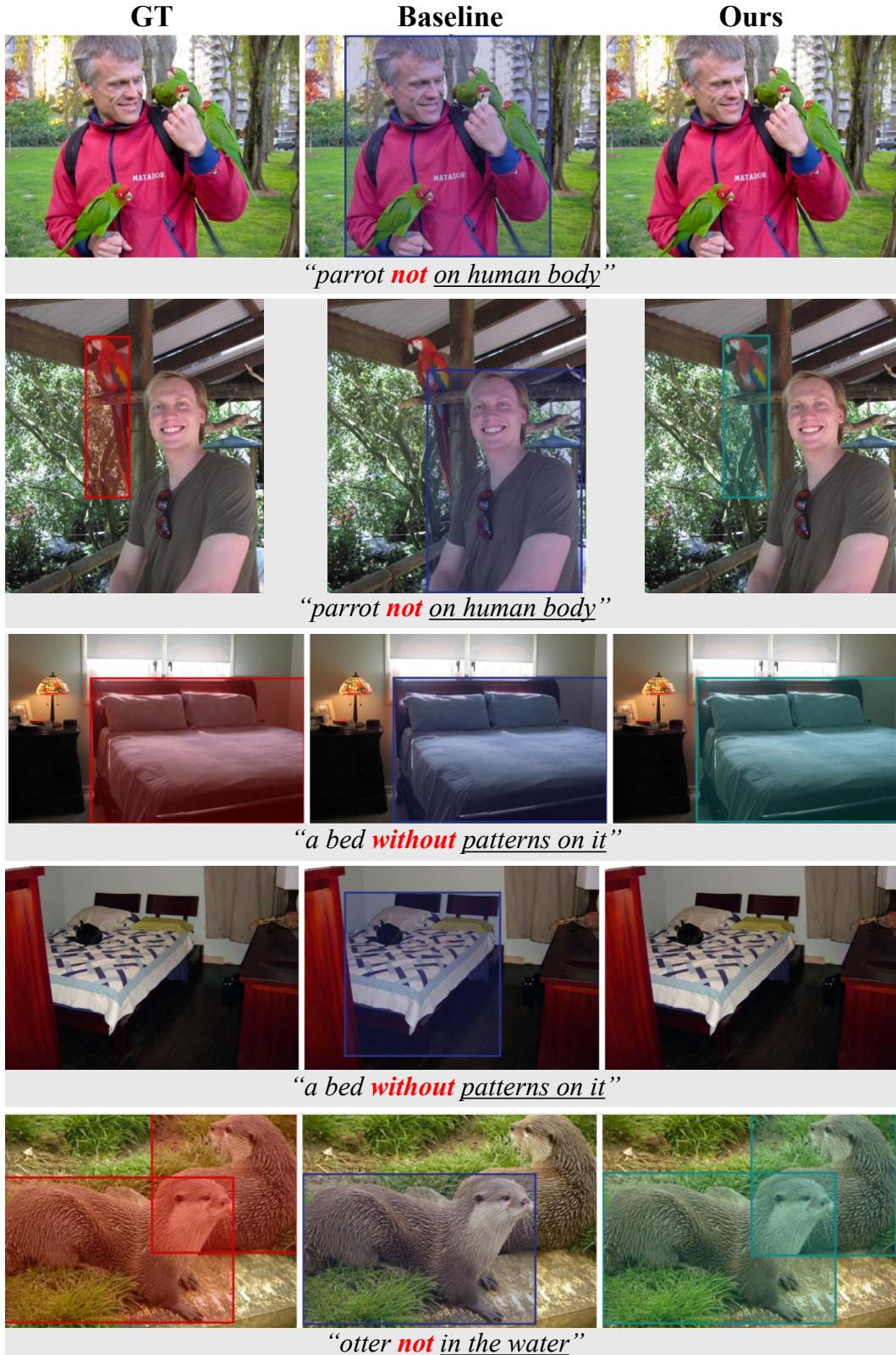
Figure S24: **Qualitative Results on D³ Datasets (3).** Absence of a bounding box means the model has determined that no instance in the image matches the input description. Our model effectively reduces the affirmative bias while keeping the correct predictions.

GT	Baseline	Ours

*"a woman in white wedding dress **not** beside any men in suits."*

*"origami **unlike** bird"*

*"a volley ball in the middle of the air **untouched** on the beach"*

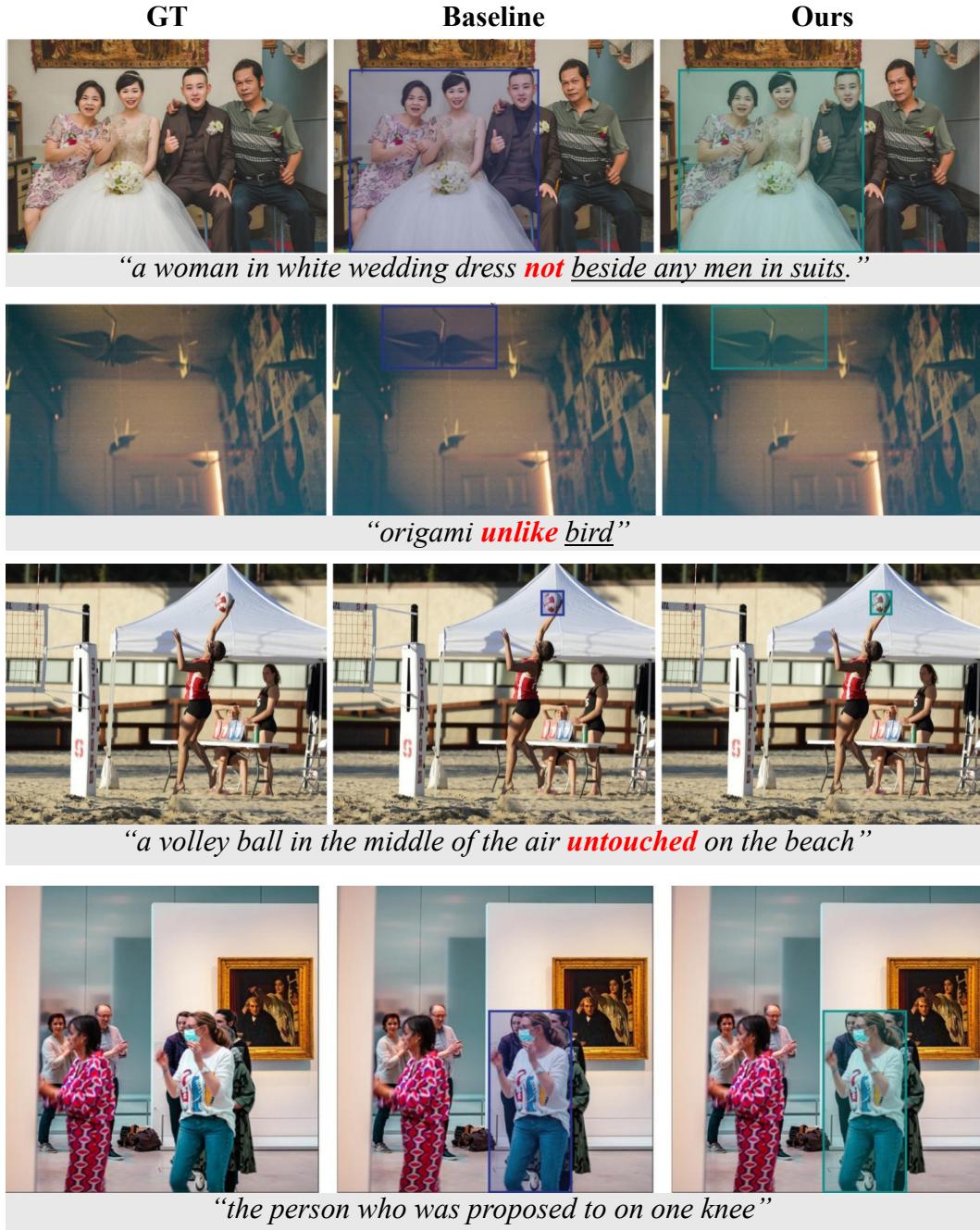*"the person who was proposed to on one knee"*

Figure S25: **Qualitative Analysis of Limitations on Complex Negation.** Despite overall improvements, our method, like the baseline, still struggles with highly complex linguistic constructions involving negation. The examples show failures in: (*i*) multi-step relational reasoning ("not beside any men in suits"), (*ii*) abstract or implicit negation ("unlike bird"), (*iii*) understanding negated states ("untouched"), and (*iv*) recognizing the absence of a complex event ("proposed to on one knee"). In these challenging cases, both models tend to default to their affirmative bias, detecting the main subject of the query rather than correctly concluding that nothing in the image matches the full description. These limitations highlight the need for more sophisticated compositional reasoning to ground complex negative constraints.