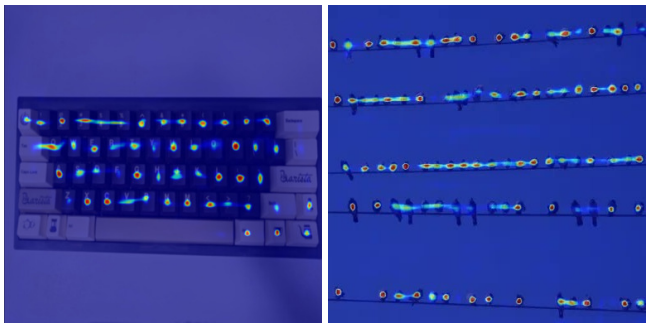
WHAT TO COUNT BEFORE HOW MANY: TEXT-RESPONSIVE ZERO-SHOT OBJECT COUNTING VIA TWO-STAGE REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-guided object counting aims to count instances specified by a language prompt without task-specific examples at inference time. Regression-based methods such as CLIP-Count Jiang et al. (2023) predict density maps efficiently, including in crowded scenes where instances overlap. However, a model may obtain a plausible count from visual repetition while responding weakly to the text prompt. Count-only metrics such as mean absolute error do not directly reveal whether changing the prompt changes which objects are counted. Detection-based methods provide explicit localization, but can incur greater computational cost in crowded scenes. We revisit density regression with a two-stage, text-guided framework. The first stage combines image and text features to predict a spatial mask for the prompted object category. The second stage uses this mask as a spatial prior to construct a prompt-conditioned similarity map and regress a density map, whose integral yields the final count. This design encourages the regression model to select objects according to the prompt without requiring instance-level detection at inference time. Experiments evaluate counting accuracy, prompt adherence, and computational cost, showing improved sensitivity to the requested category while retaining efficient inference and competitive counting performance.



(a) Text: people (b) Text: rabbit

Figure 1: Illustration of the "text-ignoring" phenomenon in existing regression-based counting methods such as ClipCount Jiang et al. (2023) and LGCount Zhang et al. (2025).

1 INTRODUCTION

Text-guided zero-shot object counting estimates the number of objects specified by a language prompt without requiring visual exemplars at inference time. The task requires a model to identify *what* matches the prompt before estimating *how many* instances are present. This distinction matters when multiple object types co-occur or when the target differs from similar objects by a fine-grained attribute.

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

Density-map regression offers an efficient way to count crowded or overlapping instances without predicting a separate bounding box for each one. Methods such as CLIP-Count Jiang et al. (2023) incorporate vision-language representations into this paradigm. Nevertheless, conditioning a regressor on text does not guarantee that its output depends on the distinguishing words in a prompt. As illustrated in Fig. 1, a model may produce a plausible count by responding to prominent repeated patterns while including objects that do not satisfy the requested description. This failure can be difficult to identify from aggregate counting error alone: mean absolute error (MAE) measures the difference between two numbers, but does not reveal which image regions produced the predicted count. Detection-based methods such as CountGD Amini-Naieni et al. (2024) make target selection explicit through localized instance predictions. Density regression instead aggregates evidence over space, which is useful for crowded scenes but can obscure whether the accumulated evidence belongs to the prompted target. The central question is therefore how to preserve density regression while making its spatial evidence responsive to the text.

We address this question with a two-stage text-guided counting framework. In the first stage, frozen vision-language encoders provide visual patch features and text-token features. Our *Text-to-Patch Relevance Gating* (T2PRG) module evaluates their fine-grained correspondence and applies channel-wise weighting to form a spatial mask. The mask indicates image regions relevant to the full prompt, including distinguishing attributes, rather than relying solely on a global text embedding. It serves as a prompt-dependent prior and is not intended to delineate individual object instances. In the second stage, we use this prior to select positive and negative visual evidence from the image. Positive evidence is drawn from regions favored by the mask, while negative evidence represents regions that should not contribute to the requested count. Comparing image features against both forms of evidence produces a contrastive similarity map. A density regressor then uses this map to predict the spatial distribution of target instances, and the integral of the density map gives the final count. In this way, the text prompt affects not only an initial feature representation but also the spatial evidence used for density estimation. We evaluate the method on FSC-147 and a diagnostic suite containing co-occurring objects and fine-grained attribute distinctions. It achieves 18.29 MAE on the diagnostic suite, compared with 39.34 for the strongest regression baseline in our comparison, and 16.37 test MAE on FSC-147.

Our contributions are summarized as follows:

- We introduce a two-stage framework that connects prompt-dependent spatial selection to density-map regression through a contrastive similarity map. This design makes the evidence used for counting explicitly dependent on the requested target.
- We propose T2PRG, which uses fine-grained text-token and image-patch interactions together with channel-wise weighting to construct a spatial prior for the prompted objects.
- We develop prompt-sensitivity evaluations based on multi-object co-occurrence and fine-grained attribute ambiguity. Holding the image fixed while varying the prompt tests target selection separately from aggregate count error.

2 Related Works

2.1 Few-shot and zero-shot object counting

Class-agnostic object counting initially evolved through exemplar-based few-shot learning. Pioneer works, including FamNet Ranjan et al. (2021) and BMNet+ Shi et al. (2022), achieved remarkable counting accuracy by matching features according to target visual exemplars. Despite their success, the requirement for manual exemplar annotations inherently hinders practical deployment, sparking a paradigm shift toward zero-shot object counting driven by textual prompts.

To enable text-driven zero-shot capabilities, recent literature heavily leverages vision-language pre-trained models (VLMs) to facilitate cross-modal grounding. Methods such as CLIP-Count Jiang et al. (2023) and LGCount Zhang et al. (2025) harness CLIP's Radford et al. (2021) joint embedding space, demonstrating impressive quantitative performance. Nevertheless, a closer inspection reveals a fundamental shortcoming: their counting predictions often decouple from the semantic nuances of user prompts, collapsing into unguided visual pattern recognition. Despite architectural efforts to enforce cross-modal alignment, current methods lack robust text-responsiveness.

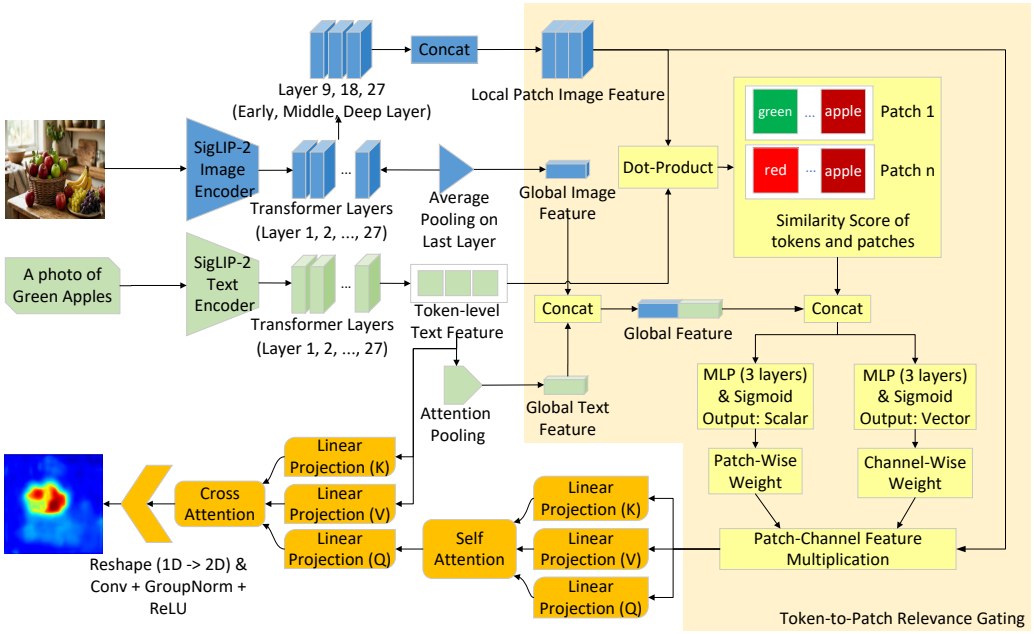


Figure 2: The overview of our architecture. This framework extracts cross-modal representations using frozen vision-language encoders. The refined features are then fused to reconstruct a text-compliant mask.

2.2 Regression-based and detection-based counting

Zero-shot object counting methods generally fall into two distinct paradigms: detection-based and regression-based. Regression-based paradigms estimate object quantities directly through continuous spatial density maps. They offer prominent advantages in computational speed, lightweight parameter scales, and memory efficiency. However, existing regression frameworks suffer from a fundamental drawback: they are structurally prone to text-ignoring, marginalizing textual constraints during count estimation and counting prompt-irrelevant visual patterns. Historically, this limitation was obscured because models could still attain acceptable MAE scores by matching visual textures.

On the other hand, Detection-based paradigms leverage open-vocabulary object detectors to explicitly locate individual instances via bounding boxes. Although detection methods overcome the text-ignoring issue and achieve high counting precision, their relies on heavy detection backbones incurs substantial computational overhead and parameter scales, rendering them less suitable for real-time or resource-constrained applications.

3 Method

3.1 Overall Pipeline

A fundamental challenge in text-guided regression-based counting is that textual supervision is only indirectly reflected in the final density-map objective. Conventional single-stage architectures jointly process the image and text and directly regress the target density map. Although cross-modal attention introduces textual information into the visual representation, the model is not explicitly required to determine whether each spatial region is relevant to the queried concept before estimating its density. Consequently, it can minimize the counting loss by exploiting salient visual patterns that correlate with object density, resulting in the text-ignoring behavior discussed in Introduction.

To explicitly constrain the counting process, we decompose text-guided counting into two stages: Text-Guided Target Localization and Mask-Guided Density Regression. Rather than allowing the density regressor to access the image and text without an explicit spatial constraint, Stage 1 first con-

structs a text-responsive mask that identifies image regions consistent with the given prompt. This mask serves as an explicit spatial bottleneck between semantic grounding and density estimation, restricting the subsequent counting process to text-relevant regions.

Stage 2 then uses the mask from stage 1 as a spatial prior to mine target and background visual exemplars. Their contrastive similarity is evaluated across multi-level image features to construct discriminative similarity maps, which are subsequently fused and regressed into the final density map. By separating semantic target identification from density estimation, the proposed framework enforces text responsiveness while preserving the computational advantages of regression-based counting.

3.2 Stage 1: Text-Guided Target Localization

The objective of Stage 1 is to determine which image regions correspond to the object described by the text before any density estimation is performed. This step is crucial because a global image-text representation alone may capture the overall semantic compatibility of an image and a prompt, but provides insufficient spatial discrimination when the prompt contains fine-grained attributes such as color, appearance, or contextual relations. Conversely, relying solely on local patch features discards the robust global image-text alignment learned by pretrained vision-language models. We therefore preserve both levels of information: local representations are used for fine-grained token-to-patch matching, while global representations provide complementary scene and query-level semantic context.

Based on this motivation, we introduce a Token-to-Patch Relevance Gating (T2PRG) module that explicitly evaluates the relevance of each visual patch to the textual prompt. Instead of directly applying cross-attention between visual and textual features, T2PRG first estimates both patch-wise and channel-wise relevance weights conditioned on local cross-modal similarity and global semantic context. The resulting gated visual representation provides an explicit prompt-dependent prior for subsequent self- and cross-attention, ultimately producing the text-compliant mask.

As shown in Figure 2, given an input image $I \in \mathbb{R}^{3\times H\times W}$, where $H = W = 384$, and a textual prompt T, we employ frozen SigLIP-2 Tschannen et al. (2025) encoders to extract visual and textual representations. Freezing the pretrained backbone preserves its open-vocabulary image-text alignment while allowing the subsequent modules to specialize in text-guided localization. For the visual branch, we extract hidden representations from layers $l \in \{9, 18, 27\}$, corresponding to early, middle, and deep semantic levels of the image encoder. These representations are concatenated to construct the local patch representation $F_{img} \in \mathbb{R}^{N\times C}$, where $N = 576$ denotes the number of image patches and $C = 1152$ denotes the feature dimension. In parallel, global average pooling over the final image representation produces the global visual embedding $G_{img} \in \mathbb{R}^{C}$. For the textual branch, the final text-encoder layer produces token-level representations $F_{txt} \in \mathbb{R}^{S\times C}$, where S denotes the text sequence length. Attention pooling over the token representations further yields a global textual representation $G_{txt} \in \mathbb{R}^{C}$. The local representations F_{img} and F_{txt} retain fine-grained patch-token correspondence, whereas G_{img} and G_{txt} encode complementary scene-level and query-level context.

We next use these representations to explicitly estimate the relevance of each image patch to the textual query. For each visual patch, we first compute its similarity to every textual token:

$$M_{dot} = \frac{F_{img}(F_{txt})^T}{\sqrt{C}}, \tag{1}$$

where M_{dot} represents fine-grained token-to-patch compatibility. Unlike a pooled text representation that compresses the entire prompt into a single vector, M_{dot} preserves the contribution of individual textual tokens and is therefore better suited for prompts containing fine-grained attributes. Token-level similarity alone, however, lacks information about the overall scene and query semantics. We therefore concatenate the global visual and textual embeddings as $G = [G_{img} \parallel G_{txt}]$, and combine G with the local image representation and token-to-patch similarity as $G' = [M_{dot} \parallel G \parallel F_{img}]$. The resulting representation jointly encodes local visual appearance, fine-grained text correspondence, and global image-text context. Two independent three-layer Multi-Layer Perceptrons(MLP) then predict a patch-wise scalar gate and a patch-specific channel-

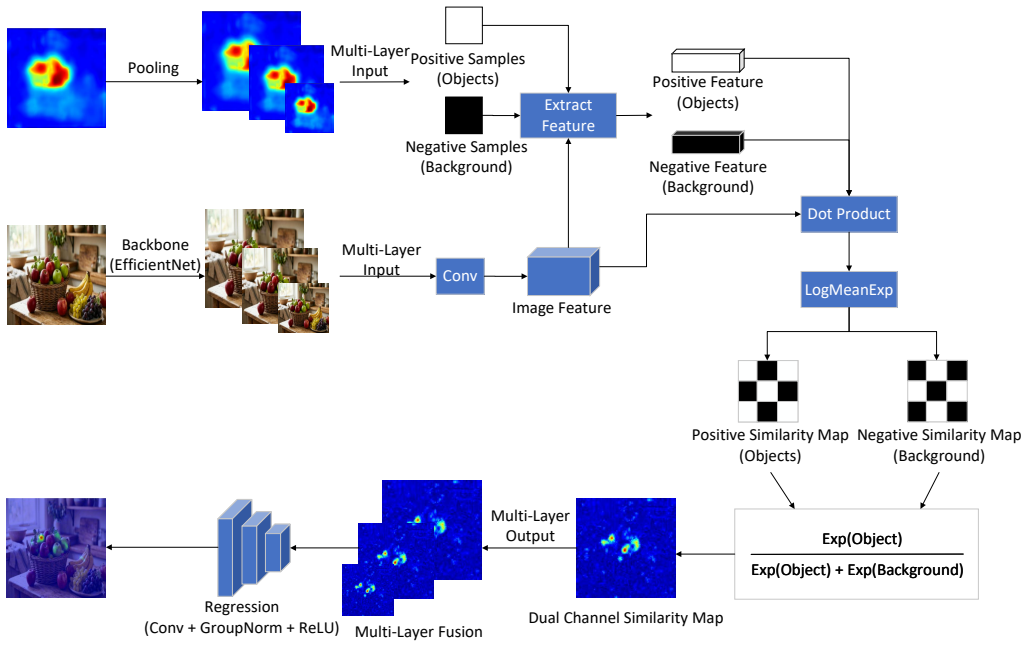


Figure 3: This stage utilizes the localization mask generated in Stage 1 as a spatial prior for feature pooling. Through dynamic exemplar mining, the model actively samples positive target and background exemplars. A dual-channel contrastive matching mechanism uses both exemplar sets to suppress ambiguous background responses.

wise gate:

$$w_{patch} = \sigma(MLP_{patch}(G')) \in [0,1]^{N\times 1}, w_{channel} = \sigma(MLP_{channel}(G')) \in [0,1]^{N\times C}, \quad (2)$$

where $\sigma(\cdot)$ denotes the sigmoid activation. The patch-wise gate determines where the text-relevant content is located, while the channel-wise gate determines which feature dimensions should be emphasized for each individual patch. The gated visual representation is subsequently obtained as:

$$X = w_{patch,i}(w_{channel,i} \odot F_{img,i}), \quad (3)$$

where $w_{patch,i} \in [0,1]$ is a scalar gate of the i-th patch, while $w_{channel,i} \in [0,1]^C$ is a channel-wise gating vector that independently modulates its C feature dimensions. The operator $\odot$ denotes element-wise multiplication. The gated representation X provides a text-conditioned prior for subsequent feature interaction. We first apply self-attention among visual patches, $X' = SelfAttention(X, X, X)$, allowing spatially separated but semantically related target regions to exchange contextual information. We then perform cross-attention using X' as the query and the token-level textual representation F_{txt} as the key and value: $X'' = CrossAttention(X', F_{txt}, F_{txt})$. This second interaction further aligns each visual representation with the textual query after irrelevant visual responses have been attenuated by T2PRG. Finally, X'' is spatially reconstructed from the patch sequence into a two-dimensional feature map and progressively refined through convolution, GroupNorm, and ReLU layers. The resulting output is a continuous mask $M \in [0,1]^{1\times H\times W}$, which explicitly represents the spatial relevance of each image region to the given textual prompt. This mask forms the semantic bottleneck between text grounding in Stage 1 and density estimation in Stage 2.

3.3 Stage 2: Mask-Guided Density Regression

Although the mask produced by Stage 1 identifies text-relevant regions, directly treating it as a counting representation is insufficient. Because M is a continuous spatial response rather than an instance-level prediction, neighboring objects in crowded scenes may produce connected high-response regions. Such mask continuity obscures instance boundaries and may retain background

regions located between closely packed targets. Therefore, Stage 2 does not directly regress the mask into a density map. Instead, we use M as a spatial prior for discovering representative target and background features. Based on this prior, we introduce Mask-Guided Dual-Channel Matching, which mines high-confidence target regions as positive exemplars and low-confidence regions as background exemplars. Rather than measuring target similarity in isolation, the model contrasts each spatial feature against both exemplar sets. This formulation converts the coarse localization prior into a discriminative object-versus-background representation and suppresses ambiguous responses introduced by mask continuity.

As illustrated in Figure 3, a frozen EfficientNet Tan & Le (2019) backbone extracts visual representations $F_i \in \mathbb{R}^{C_i \times H_i \times W_i}$, $i \in \{1, 2, 3\}$, where $C_1 = 32, C_2 = 48, C_3 = 136$, and the corresponding spatial resolutions are $(H_1, W_1) = (96, 96), (H_2, W_2) = (48, 48)$, and $(H_3, W_3) = (24, 24)$, respectively. Using multiple feature levels allows the matching mechanism to capture target evidence under different receptive fields. The mask M is adaptively max-pooled to each spatial resolution, yielding scale-specific priors M_i. At each level, we construct two exemplar sets:

- Positive Exemplars (Q_{obj}): features sampled from regions where $M_i > \tau_{obj}$, with $\tau_{obj} = 0.8$, representing text-specific target regions.
- Negative Exemplars (Q_{bg}): features sampled from regions where $M_i < \tau_{bg}$, with $\tau_{bg} = 0.5$, representing background regions.

By mining both sets from the current image rather than relying on fixed category prototypes, the resulting exemplars adapt to the appearance and background context of each input.

Since multiple target or background exemplars may be extracted from the same image, relying on a single exemplar can produce unstable matching responses. We therefore compare every spatial feature against all mined exemplars and aggregate their matching evidence using LogMeanExp:

$$S_{obj}(x, y) = \log \left(\frac{1}{|Q_{obj}|} \sum_{q \in Q_{obj}} \exp \left(\frac{q^T F_i(x, y)}{\sqrt{C}} \right) \right), \tag{4}$$

with S_{bg} computed analogously from Q_{bg}. Nevertheless, a high object similarity alone does not necessarily indicate a reliable target response, because visually ambiguous background regions may also exhibit high similarity. We therefore normalize the object response against its competing background response:

$$S_i = \frac{\exp(S_{obj})}{\exp(S_{obj}) + \exp(S_{bg})}. \tag{5}$$

This contrastive formulation produces an object-probability map conditioned on both positive and negative evidence. In particular, spatial regions that are simultaneously similar to target and background exemplars are suppressed, which helps separate adjacent targets and remove residual clutter retained by the Stage 1 mask.

The contrastive similarity maps obtained from different feature levels capture complementary target evidence under different receptive fields. We therefore resize them to a common spatial resolution of 96×96 and fuse them along the channel dimension using a 1×1 convolution. The fused representation is then forwarded to a progressive regression head composed of upsampling, convolution, GroupNorm, and ReLU layers to produce the final density map $D \in \mathbb{R}^{1 \times 384 \times 384}$. The final object count is obtained by integrating the predicted density over its spatial dimensions: $Count = \sum_{x=1}^{H} \sum_{y=1}^{W} D(x, y)$.

3.4 Optimization Objectives

Stage 1 Optimization Objectives. Stage 1 is optimized to generate a continuous localization mask that spatially corresponds to the objects specified by the text. Given the predicted mask M and ground-truth binary mask $\hat{M}$, we employ binary cross-entropy:

$$L_{BCE} = \hat{M} \log(M) + (1 - \hat{M}) \log(1 - M), \tag{6}$$

Table 1: Quantitative comparison on our targeted diagnostic suite. Our method achieves superior prompt responsiveness over existing regression models.

Method	Venue	Paradigm	Evaluations	
			MAE ↓	RMSE ↓
CountGD Amini-Naieni et al. (2024)	NeurIPS'24	Detection	11.10	47.21
CAD-GD Wang et al. (2025b)	CVPR'25	Detection	9.63	41.28
CountSE Liu et al. (2025)	ICCV'25	Detection	8.44	52.57
FamNet Ranjan et al. (2021)	CVPR'21	Regression	42.54	87.11
ZSC Xu et al. (2023)	CVPR'23	Regression	58.44	95.47
ClipCount Jiang et al. (2023)	MM'23	Regression	41.57	74.15
VLCounter Kang et al. (2024)	AAAI'24	Regression	39.63	76.28
CounTX Amini-Naieni et al. (2023)	BMVC'23	Regression	39.34	77.29
VACount Zhu et al. (2024)	ECCV'24	Regression	51.22	80.67
LGCount Zhang et al. (2025)	ICCV'25	Regression	42.98	77.72
Ours	–	Regression	18.29	33.64

Stage 2 Density Regression Objectives. Stage 2 is supervised at both the final density-estimation level and the intermediate similarity-matching level. The density regression objective minimizes the Mean Squared Error(MSE) between the predicted density map D and ground-truth density map $\hat{D}$: $L_{MSE} = \left(D - \hat{D}\right)^2$. In addition, we supervise the intermediate contrastive similarity maps to encourage separation between target and background responses:

$$L_{con} = mean\left(\frac{1}{2}\left(L_{BCE}(\Delta s, P) + L_{BCE}(S_{obj}, P)\right)\right), \tag{7}$$

where $\Delta s_i = s_{obj} - s_{bg}$ denotes the contrastive logit difference, S_i denotes the object probability defined in Eq. (5), and P^i denotes the corresponding binary supervision at feature level i. The supervision masks are constructed by adaptively pooling the ground-truth density map and thresholding responses above 20% of its maximum value. The overall loss of stage 2 is formulated as $L_{MSE} + L_{con}$.

4 Experiments

4.1 Experimental Setup

Text-Responsiveness Diagnostic Suite. Standard zero-shot counting benchmarks are dominated by images containing a single target category, making it difficult to determine whether a model truly follows textual constraints or simply responds to visually salient object patterns. To explicitly evaluate prompt responsiveness, we construct a diagnostic suite containing 60 multi-object real-world scenes with fine-grained attribute ambiguities, including objects from the same category with different colors or subcategories. Paired with contrasting textual prompts, the suite contains 150 image-text pairs and directly evaluates whether counting predictions change consistently with the specified target.

Standard Zero-Shot Counting. We evaluate conventional counting accuracy on FSC-147 dataset, which contains 6,135 images across 147 object categories. Following the standard zero-shot protocol, we report results on the official validation and test splits using textual prompts without visual exemplars.

Computational Efficiency. We additionally evaluate model complexity and inference throughput using a single NVIDIA RTX 3090Ti with batch size 1 and an input resolution of 384×384. We report the total number of model parameters (Params) and end-to-end inference throughput in frames per second (FPS).

Table 2: Comparison on the FSC-147 benchmark under the zero-shot setting. Our framework maintains competitive counting accuracy among regression-based approaches while explicitly enforcing textual responsiveness.

Method	Venue	Paradigm	Val Set		Test Set	
			MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
CountGD [1]	NeurIPS'24	Detection	12.14	47.51	14.76	140.52
CAD-GD [2]	CVPR'25	Detection	9.30	40.96	10.35	86.88
CountSE [3]	ICCV'25	Detection	8.51	54.93	7.84	82.99
FamNet [4]	CVPR'21	Regression	21.55	67.08	20.61	111.54
ZSC [5]	CVPR'23	Regression	26.93	88.63	22.09	115.17
ClipCount [6]	MM'23	Regression	18.93	66.10	17.45	104.88
VLCounter [7]	AAAI'24	Regression	18.06	65.13	17.05	106.16
CounTX [8]	BMVC'23	Regression	17.81	63.61	16.95	102.43
VACount [9]	ECCV'24	Regression	18.11	73.22	17.88	110.31
LGCount [10]	ICCV'25	Regression	17.52	58.39	16.42	100.63
Ours	–	Regression	18.08	65.47	16.37	104.48

Note: Citations in the table correspond to [1]Amini-Naieni et al. (2024) [2]Wang et al. (2025b) [3]Liu et al. (2025) [4]Ranjan et al. (2021) [5]Xu et al. (2023) [6]Jiang et al. (2023) [7]Kang et al. (2024) [8]Amini-Naieni et al. (2023) [9]Zhu et al. (2024) [10]Zhang et al. (2025)

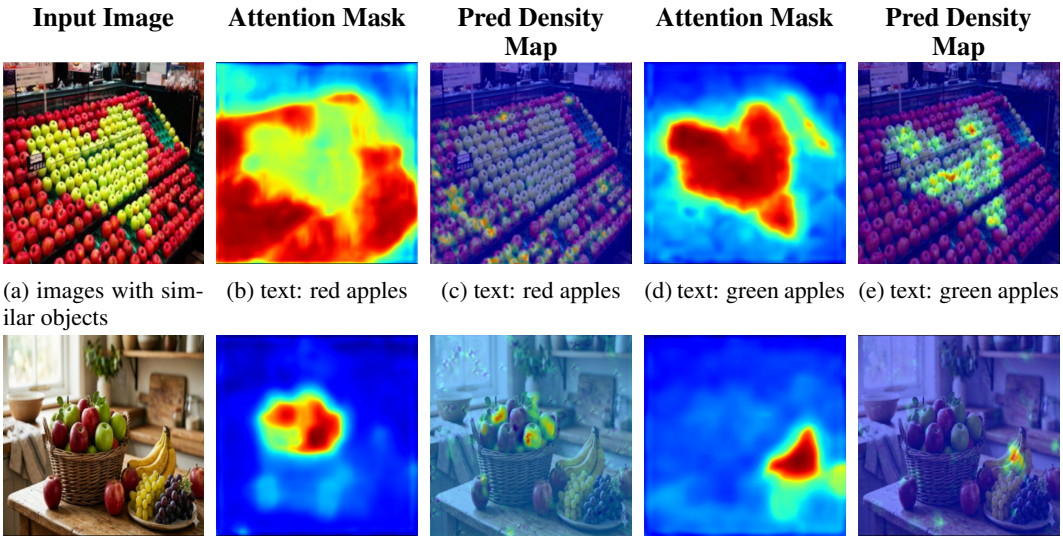
Table 3: Inference efficiency and computational cost comparison on a single NVIDIA RTX 3090Ti.

Method	Paradigm	Params ↓	FPS ↑
CountGD Amini-Naieni et al. (2024)	Detection	146 M	4.93
CAD-GD Wang et al. (2025b)	Detection	219 M	3.24
CountSE Liu et al. (2025)	Detection	135 M	5.78
CLIP-Count Jiang et al. (2023)	Regression	101 M	10.34
VLCounter Kang et al. (2024)	Regression	103 M	8.45
CounTX Amini-Naieni et al. (2023)	Regression	97 M	9.98
VACount Zhu et al. (2024)	Regression	108 M	9.10
LGCount Zhang et al. (2025)	Regression	110 M	10.07
Ours	Regression	92 M	10.79

For counting accuracy, we follow standard class-agnostic counting protocols and report Mean Absolute Error (MAE): $\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i|$, and Root Mean Squared Error (RMSE): $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}$, where N denotes the total number of evaluation images, y_i is the predicted count, and $\hat{y}_i$ is the corresponding ground truth count. Lower values of MAE and RMSE indicate better counting accuracy.

4.2 Results

Text Responsiveness. Table 1 reports results on our diagnostic suite. Existing regression-based approaches exhibit substantially higher counting errors, with the strongest competing regression method achieving an MAE of 39.34. In comparison, our method reduces the MAE to 18.29, corresponding to a 53.5% relative reduction. This large improvement indicates that explicitly separating text-guided localization from density estimation substantially improves robustness to fine-grained prompt changes. Compared with detection-based approaches, our method still exhibits a higher MAE, but achieves the lowest RMSE among all evaluated methods. These results suggest that the proposed regression framework substantially narrows the robustness gap between regression- and detection-based paradigms while retaining a lightweight counting architecture.



(a) images with similar objects (b) text: red apples (c) text: red apples (d) text: green apples (e) text: green apples

(f) image with various objects (g) text: green apples in basket (h) text: green apples in basket (i) text: bananas on table (j) text: bananas on table

Figure 4: Qualitative results on multi-object co-occurrence scenes. Given the same image, different textual prompts produce distinct masks and corresponding density maps, demonstrating text-dependent counting behavior.

Standard Zero-Shot Counting. Table 2 evaluates counting accuracy on the standard FSC-147 benchmark. Our model achieves competitive performance among regression-based methods, with a Test MAE of 16.37, slightly outperforming the other regression approaches considered in our comparison. Although our validation performance and RMSE do not consistently surpass the strongest regression baselines, the results show that introducing explicit text-guided localization does not substantially compromise conventional counting accuracy. Detection-based methods achieve lower MAE on this benchmark, but rely on considerably larger detection-oriented architectures. Together with the diagnostic results in Table 1, these results demonstrate that our framework improves prompt responsiveness while preserving the counting capability of lightweight regression-based models.

Computational Efficiency. Table 3 compares model size and inference throughput. Our framework contains 92M parameters and achieves 10.79 FPS, corresponding to both the smallest parameter footprint and the highest throughput among the compared methods. In contrast, the evaluated detection-based approaches require 135–219M parameters and operate at 3.24–5.78 FPS. These results show that explicit text-guided localization can be incorporated into a regression framework without sacrificing its computational efficiency.

5 Conclusion

In this work, we revisit regression-based zero-shot object counting from the perspective of text responsiveness. We identify that conventional regression frameworks can produce plausible counting results while relying excessively on visual patterns and insufficiently responding to fine-grained textual constraints. To address this limitation, we propose a two-stage framework that explicitly separates text-guided target localization from density estimation, thereby constraining the counting process with a text-compliant spatial prior. Experiments on our targeted diagnostic suite demonstrate substantially improved prompt responsiveness in challenging multi-object and attribute-ambiguous scenes. Meanwhile, the proposed framework maintains competitive counting accuracy on the standard FSC-147 benchmark and preserves the computational efficiency of regression-based approaches. These results suggest that robust text-guided counting does not necessarily require heavyweight detection-based architectures, and that explicitly constraining what to count provides a promising direction for efficient zero-shot object counting.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, grammatical correction, improving the clarity and organization of manuscript text. Generative AI was not used to generate experimental results, fabricate data, or independently conduct the experiments reported in this work. All AI-assisted text and suggestions were manually reviewed, revised, and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Niki Amini-Naieni and Andrew Zisserman. Countgd++: Generalized prompting for open-world counting. *arXiv preprint arXiv:2512.23351*, 2025.

Niki Amini-Naieni and Andrew Zisserman. Open-world object counting in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2300–2308, 2026.

Niki Amini-Naieni, Kiana Amini-Naieni, Tengda Han, and Andrew Zisserman. Open-world text-specified object counting. *arXiv preprint arXiv:2306.01851*, 2023.

Niki Amini-Naieni, Tengda Han, and Andrew Zisserman. Countgd: Multi-modal open-world counting. *Advances in Neural Information Processing Systems*, 37:48810–48837, 2024.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Jian Cheng, Haipeng Xiong, Zhiguo Cao, and Hao Lu. Decoupled two-stage crowd counting and beyond. *IEEE Transactions on Image Processing*, 30:2862–2875, 2021.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16901–16911, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, pp. 4145–4153, 2017.

Ruixiang Jiang, Lingbo Liu, and Changwen Chen. Clip-count: Towards text-guided zero-shot object counting. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4535–4545, 2023.

Seunggu Kang, WonJun Moon, Euiyeon Kim, and Jae-Pil Heo. Vlcounter: Text-aware visual representation for zero-shot object counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2714–2722, 2024.

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.

Shuai Liu, Peng Zhang, Shiwei Zhang, and Wei Ke. Countse: Soft exemplar open-set object counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21536–21546, 2025.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7086–7096, 2022.

Andreas Michel, Wolfgang Gross, Fabian Schenkel, and Wolfgang Middelmann. Class-aware object counting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 469–478, 2022.

Jer Pelhan, Vitjan Zavrtanik, Matej Kristan, et al. Dave-a detect-and-verify paradigm for low-shot counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23293–23302, 2024.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yifei Qian, Zhongliang Guo, Bowen Deng, Chun Tong Lei, Shuai Zhao, Chun Pong Lau, Xiaopeng Hong, and Michael P Pound. T2icount: Enhancing cross-modal understanding for zero-shot counting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25336–25345, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3394–3403, 2021.

Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and Zhiguo Cao. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9529–9538, 2022.

Zenglin Shi, Ying Sun, and Mengmi Zhang. Training-free object counting with prompts. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 323–331, 2024.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PmLR, 2019.

Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. *Advances in Neural Information Processing Systems*, 36:46830–46855, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Bo Wan, Michael Tschannen, Yongqin Xian, Filip Pavetic, Ibrahim Alabdulmohsin, Xiao Wang, André S Pinto, Andreas Steiner, Lucas Beyer, and Xiaohua Zhai. Locca: Visual pretraining with location-aware captioners. *Advances in Neural Information Processing Systems*, 37:116355–116387, 2024.

Yaxian Wang, Henghui Ding, Shuting He, Xudong Jiang, Bifan Wei, and Jun Liu. Hierarchical alignment-enhanced adaptive grounding network for generalized referring expression comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8042–8050, 2025a.

Zhicheng Wang, Zhiyu Pan, Zhan Peng, Jian Cheng, Liwen Xiao, Wei Jiang, and Zhiguo Cao. Exploring contextual attribute density in referring expression counting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19587–19596, 2025b.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10216–10225, 2020.

Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. Zero-shot object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15548–15557, 2023.

Guanning Zeng, Xiang Zhang, Zirui Wang, Haiyang Xu, Zeyuan Chen, Bingnan Li, and Zhuowen Tu. Yolo-count: Differentiable object counting for text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16765–16775, 2025.

Chao Zhai, Fan Zhang, Wenyu Chen, Malu Zhang, Fan Li, and Xuanting Xie. Decoupled feature matching for few-shot counting and localization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 833–841, 2015.

Shiwei Zhang, Qi Zhou, and Wei Ke. Enhancing zero-shot object counting via text-guided local ranking and number-evoked global attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21097–21106, 2025.

Zixuan Zheng, Yilei Shi, Chunlei Li, Jingliang Hu, Xiao Xiang Zhu, and Lichao Mou. Rethinking cell counting methods: Decoupling counting and localization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 418–426. Springer, 2024.

Huilin Zhu, Jingling Yuan, Zhengwei Yang, Yu Guo, Zheng Wang, Xian Zhong, and Shengfeng He. Zero-shot object counting with good exemplars. In *European conference on computer vision*, pp. 368–385. Springer, 2024.

Yuda Zou, Zijian Zhang, and Yongchao Xu. Decoupling what to count and where to see for referring expression counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 14113–14121, 2026.

A Appendix

A.1 Implementation Details

All experiments are implemented in PyTorch and conducted on a single NVIDIA RTX 3090Ti GPU. The pretrained visual backbones remain frozen throughout training. For Stage 1, we adopt SigLIP-2 base-patch16-384 as the vision-language backbone and train the proposed text-guided localization modules on PhraseCut dataset for 20 epochs. We use AdamW with a fixed learning rate of $1 \times 10^{-5}$and a weight decay of 1×10^{-4}. For Stage 2, EfficientNet-B3 is employed as the frozen visual backbone. The mask-guided density regression modules are trained on FSC-147 dataset for 100 epochs using AdamW with a fixed learning rate of 1×10^{-5}.

A.2 Cross-Dataset Generalization

To further evaluate the robustness of the proposed framework beyond the standard FSC-147 benchmark, we conduct additional experiments on CARPK and ShanghaiTech. These datasets exhibit substantially different scene characteristics, object distributions, and density levels from FSC-147, providing a complementary evaluation of the generalization capability of our framework. We report MAE and RMSE following the same evaluation criteria used in the main paper.

Generalization to CARPK dataset. As shown in Table 4, our method achieves an MAE of 7.81 and an RMSE of 9.92 on CARPK. Among regression-based approaches, our framework remains competitive with recent methods, while detection-based approaches achieve stronger absolute counting performance, our results demonstrate that the proposed regression framework retains competitive counting capability under a substantially different evaluation domain.

Table 4: Quantitative comparison on CARPK dataset.

Method	Venue	Paradigm	Evaluations	
			MAE ↓	RMSE ↓
CountGD Amini-Naieni et al. (2024)	NeurIPS'24	Detection	3.93	5.41
CAD-GD Wang et al. (2025b)	CVPR'25	Detection	3.14	5.36
CountSE Liu et al. (2025)	ICCV'25	Detection	2.79	4.20
FamNet Ranjan et al. (2021)	CVPR'21	Regression	18.19	33.66
BMNet+ Shi et al. (2022)	CVPR'22	Regression	10.44	13.77
ClipCount Jiang et al. (2023)	MM'23	Regression	11.96	16.61
VLCounter Kang et al. (2024)	AAAI'24	Regression	6.46	8.68
CounTX Amini-Naieni et al. (2023)	BMVC'23	Regression	8.13	10.87
VACount Zhu et al. (2024)	ECCV'24	Regression	10.63	13.20
LGCount Zhang et al. (2025)	ICCV'25	Regression	7.79	9.21
Ours	–	Regression	7.81	9.92

Table 5: Comparison on the ShanghaiTech dataset.

Method	Venue	Paradigm	Part A		Part B	
			MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
CountGD [1]	NeurIPS'24	Detection	116.02	234.07	28.16	50.44
CAD-GD [2]	CVPR'25	Detection	132.21	253.93	32.16	57.87
CountSE [3]	ICCV'25	Detection	129.74	258.25	30.78	55.65
ClipCount [4]	MM'23	Regression	192.62	308.37	45.68	77.43
VLCounter [5]	AAAI'24	Regression	186.87	292.22	40.34	74.36
CounTX [6]	BMVC'23	Regression	178.14	278.24	38.05	67.48
VACount [7]	ECCV'24	Regression	197.43	302.90	48.66	76.89
LGCount [8]	ICCV'25	Regression	169.92	284.38	41.17	73.67
Ours	–	Regression	183.55	285.96	39.27	71.35

Note: Citations in the table correspond to [1]Amini-Naieni et al. (2024) [2]Wang et al. (2025b) [3]Liu et al. (2025) [4]Jiang et al. (2023) [5]Kang et al. (2024) [6]Amini-Naieni et al. (2023) [7]Zhu et al. (2024) [8]Zhang et al. (2025)

Generalization to ShanghaiTech dataset. We further evaluate the proposed framework on the ShanghaiTech benchmark, which contains substantially denser counting scenarios. As reported in Table 5, our method obtains an MAE/RMSE of 183.55/285.96 on Part A and 39.27/71.35 on Part B. These results further suggest that explicitly separating text-guided localization from density estimation does not restrict the framework to the distribution of FSC-147 and allows it to maintain competitive counting performance across substantially different scene densities.

B Ablation Study

To systematically evaluate how each core design addresses the text-ignoring issue and density estimation limitations, we conduct ablation studies across both the Multi-Object diagnostic suite and the FSC-147 dataset. The overall ablation suite consists of three targeted comparative setups designed to isolate: (1) the patch-channel gating mechanism in Stage 1, (2) the choice of vision-language backbones, and (3) the dual-channel background suppression mechanism in Stage 2.

Effectiveness of the Dual Spatial-Channel Gating (T2PRG) Module. Counting models, including regression-based and detection-based frequently, suffer from complex attribute prompts. To

Table 6: Ablation study of our proposed method with and without T2PRG module during stage 1, evaluated on the Multi-Object diagnostic suite and FSC-147 dataset.

Variables	Multi-Object		FSC-147 (Val)		FSC-147 (Test)	
	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
Full T2PRG Module	18.29	33.64	18.08	65.47	16.37	104.48
Patch-wise only	19.25	36.69	19.11	65.62	16.53	105.85
Channel-wise only	20.66	41.03	19.54	67.42	16.79	105.92
Without T2PRG Module	22.14	47.39	19.73	68.87	16.91	110.05

determine whether our proposed T2PRG module successfully forces visual features to comply with fine-grained textual modifiers, we construct a comparative baseline by removing the T2PRG module (*w/o T2PRG*) from Stage 1 while keeping all other architecture parameters identical.

Influence of Vision-Language Backbone. A central question in vision-language frameworks is whether cross-modal alignment performance stems from explicit architectural design or simply from leveraging stronger pretrained vision-language encoders. To evaluate the architectural dependence on backbone representations, we set up an experiment replacing the default SigLIP-2 encoder with a standard CLIP (ViT-B/16) backbone while retaining the remaining pipeline.

Impact of Dual-Channel Matching. Localization masks generated in Stage 1 inherently suffer from mask continuity, where spatially adjacent objects in congested scenes merge into continuous blobs and trap background clutter between targets. To verify whether explicitly mining and subtracting negative background exemplars resolves this spatial ambiguity, we compare our dual-channel contrastive matcher ($S_i = \frac{\exp(S_{obj})}{\exp(S_{obj}) + \exp(S_{bg})}$) against a single-channel baseline ($S_i = \frac{\exp(S_{obj})}{\exp(S_{obj}) + \exp(0)}$) that relies exclusively on positive target matching.

B.1 Results

The quantitative evaluations across all ablation variants are summarized in Tables 6–8, with qualitative heatmap and density predictions presented in Figures 5 and 6.

As indicated by the performance metrics in Table 1, removing the T2PRG module leads to noticeable error inflation across both evaluation benchmarks. The visual evidence in Figure 1(c, g, k) confirms that without spatial-channel gating, the model loses target confidence and exhibits diffuse activations over prompt-irrelevant objects (e.g., falsely responding to red apples under a green apple prompt, or activating non-bronze coins). These observations confirm that channel-wise denoising paired with token-level spatial gating is essential for constraining attention strictly to specified textual attributes.

Regarding backbone selection, Table 2 and Figure 1(d, h, l) show that replacing SigLIP-2 with CLIP results in reduced spatial stability and higher overall error rates. This divergence reflects their underlying pre-training objectives: while SigLIP-2 incorporates dense patch-level features well-suited for localized downstream alignment, standard CLIP is optimized primarily for global image-text correspondence, rendering its representations less stable for fine-grained spatial localization.

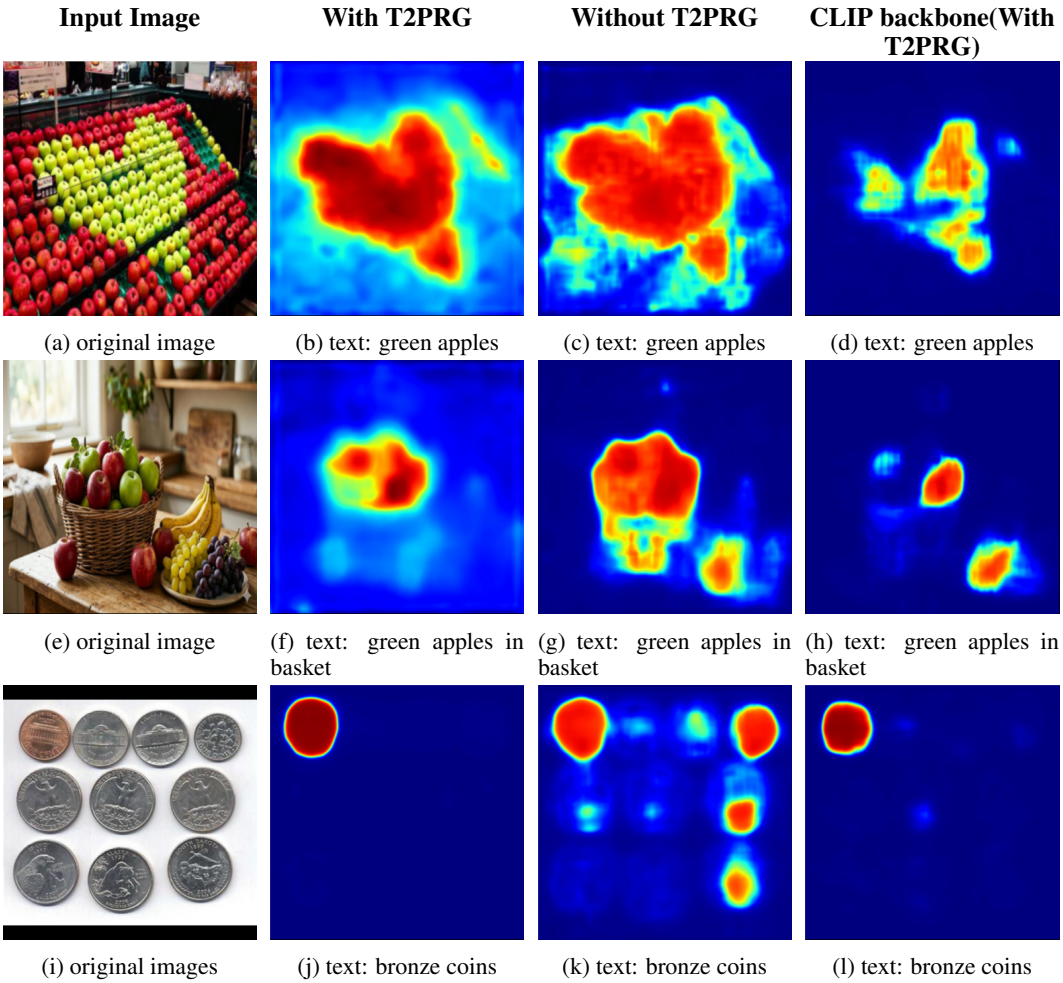
The quantitative results in Table 3 demonstrate a substantial performance gap between the matching strategies in Stage 2, with single-channel matching yielding significantly higher counting errors across all datasets. As visually illustrated in Figure 2(d, h, l), relying solely on positive target similarity (S_{obj}) causes broad false-positive activations across non-target regions, as positive matching alone cannot separate targets from the mixed background context present in continuous probability maps, leading to difficulty during regression training. In contrast, our dual-channel formulation actively utilizes negative background logits (S_{bg}), effectively canceling out lingering background noise trapped between closely packed instances and sharpening target boundaries for density regression.

Table 7: Ablation study of our proposed method with pretrained SigLIP2 and CLIP backbone during stage 1, evaluated on the Multi-Object diagnostic suite and FSC-147 dataset.

Backbone	Multi-Object		FSC-147 (Val)		FSC-147 (Test)	
	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
SigLIP-2	18.29	33.64	18.08	65.47	16.37	104.48
CLIP	19.78	36.28	18.23	65.79	17.60	108.11

Table 8: Ablation study of our proposed method with single and dual channel during stage 2, evaluated on the Multi-Object diagnostic suite and FSC-147 dataset.

Variables	Multi-Object		FSC-147 (Val)		FSC-147 (Test)	
	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
Dual Channel (S_{obj}&S_{bg})	18.29	33.64	18.08	65.47	16.37	104.48
Single Channel (S_{obj})	27.10	51.22	26.84	92.78	25.13	144.67

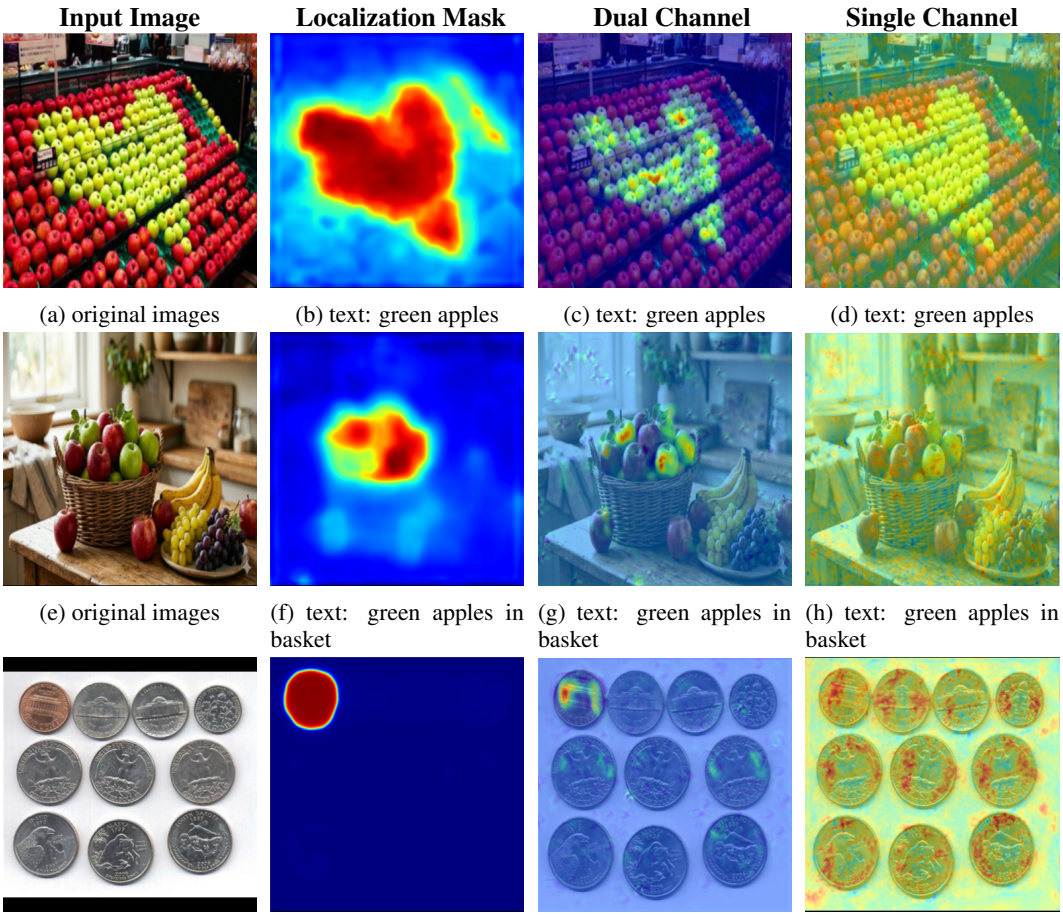


(a) original image (b) text: green apples (c) text: green apples (d) text: green apples

(e) original image (f) text: green apples in basket (g) text: green apples in basket (h) text: green apples in basket

(i) original images (j) text: bronze coins (k) text: bronze coins (l) text: bronze coins

Figure 5: Qualitative comparison of Stage 1 model variants. (a, e, i): Input images, (b, f, j): Localization masks generated by our full model (SigLIP-2 with T2PRG). (c, g, k): Masks generated by the baseline without the T2PRG module. (d, h, l): Masks generated by replacing SigLIP-2 with a standard CLIP backbone.



(a) original images (b) text: green apples (c) text: green apples (d) text: green apples

(e) original images (f) text: green apples in basket (g) text: green apples in basket (h) text: green apples in basket

(i) original image (j) text: bronze coins (k) text: bronze coins (l) text: bronze coins

Figure 6: Qualitative comparison of Stage 2 matching strategies. (a, e, i): Input images. (b, f, j): Localization masks from Stage 1, serving as spatial priors. (c, g, k): Final density maps predicted by our proposed dual-channel matcher (S_{obj}&S_{bg}). (d, h, l): Density maps predicted by the single-channel baseline relying only on positive targets (S_{obj}).

B.2 ALGORITHMS

For completeness, Algorithms 1 and 2 provide the detailed forward procedures of the two stages in our framework. Algorithm 1 summarizes text-guided target localization. Given an image and a textual prompt, frozen vision-language encoders first extract local and global cross-modal representations. T2PRG subsequently estimates patch-wise and channel-wise relevance weights, followed by self and cross-attention refinement to produce the text-responsive mask.

Algorithm 1 Stage 1: Text-Guided Target Localization

Require: Image I, text prompt T
Ensure: Mask M

$\{F_{img}^{(l)}\}_{l=1}^{27} = ImageEncoder(I)$;
$F_{txt} = TextEncoder(T)$;
$F_{img} = Concat(F_{img}^{(9)}, F_{img}^{(18)}, F_{img}^{(27)})$;
$G_{img} = AveragePool(F_{img}^{(27)})$;
$G_{txt} = AttentionPool(F_{txt})$;
$M_{dot} = \frac{F_{img} F_{txt}^{T}}{\sqrt{C}}$;
$G = Concat(G_{img}, G_{txt})$;
$G' = Concat(M_{dot}, G, F_{img})$;
$w_{patch} = \sigma(MLP_{patch}(G'))$;
$w_{channel} = \sigma(MLP_{channel}(G'))$;
$X = w_{patch} w_{channel} F_{img}$;
$X' = SelfAttention(X, X, X)$;
$X'' = CrossAttention(X', F_{txt}, F_{txt})$;
$M = ReLU(GroupNorm(Conv(Reshape(X''))))$;
$M = \sigma(M)$;
return M

Algorithm 2 describes mask-guided density regression. Rather than directly converting the Stage 1 mask into a density estimate, the mask is used as a spatial prior to mine high-confidence target and background exemplars at multiple feature levels. Their contrastive matching responses are aggregated into multi-scale similarity maps, which are subsequently fused and transformed into the final density map. The predicted object count is obtained by integrating the density map over its spatial dimensions.

Algorithm 2 Stage 2: Mask-Guided Density Regression

Require: Image I, mask M, thresholds τ_{obj}, τ_{bg}
Ensure: Density map D, predicted count estimation $Count$

$\{F_i\}_{i=1}^{3} = \text{EfficientNet}(I)$;
for $i = 1$ to 3 **do**
    $M_i = AdaptiveMaxPool(M, H_i, W_i)$;
    $Q_{obj} = \{F_i(x, y) \mid M_i(x, y) > \tau_{obj}\}$;
    $Q_{bg} = \{F_i(x, y) \mid M_i(x, y) < \tau_{bg}\}$;
    for all $(x, y) \in F_i$ **do**
        $S^i_{obj}(x, y) = LogMeanExp\left(\left\{\frac{q^T F_i(x,y)}{\sqrt{C_i}} \mid q \in Q_{obj}\right\}\right)$;
        $S^i_{bg}(x, y) = LogMeanExp\left(\left\{\frac{q^T F_i(x,y)}{\sqrt{C_i}} \mid q \in Q_{bg}\right\}\right)$;
    end for
    $S_i = \dfrac{\exp(S^i_{\text{obj}})}{\exp(S^i_{\text{obj}}) + \exp(S^i_{\text{bg}})}$;
end for
$Fusion = Conv\left(Concat(S_1, S_2, S_3)\right)$;
$D = ReLU(GroupNorm(Conv(Upsample(Fusion))))$;
$Count = \sum_{x=1}^{H} \sum_{y=1}^{W} D(x, y)$;
return $D, Count$