This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

YOLO-Count: Differentiable Object Counting for Text-to-Image Generation

Guanning Zeng[1*] Xiang Zhang[2] Zirui Wang[3] Haiyang Xu[2]
Zeyuan Chen[2] Bingnan Li[2] Zhuowen Tu[2]
[1]Tsinghua University [2]UC San Diego [3]UC Berkeley

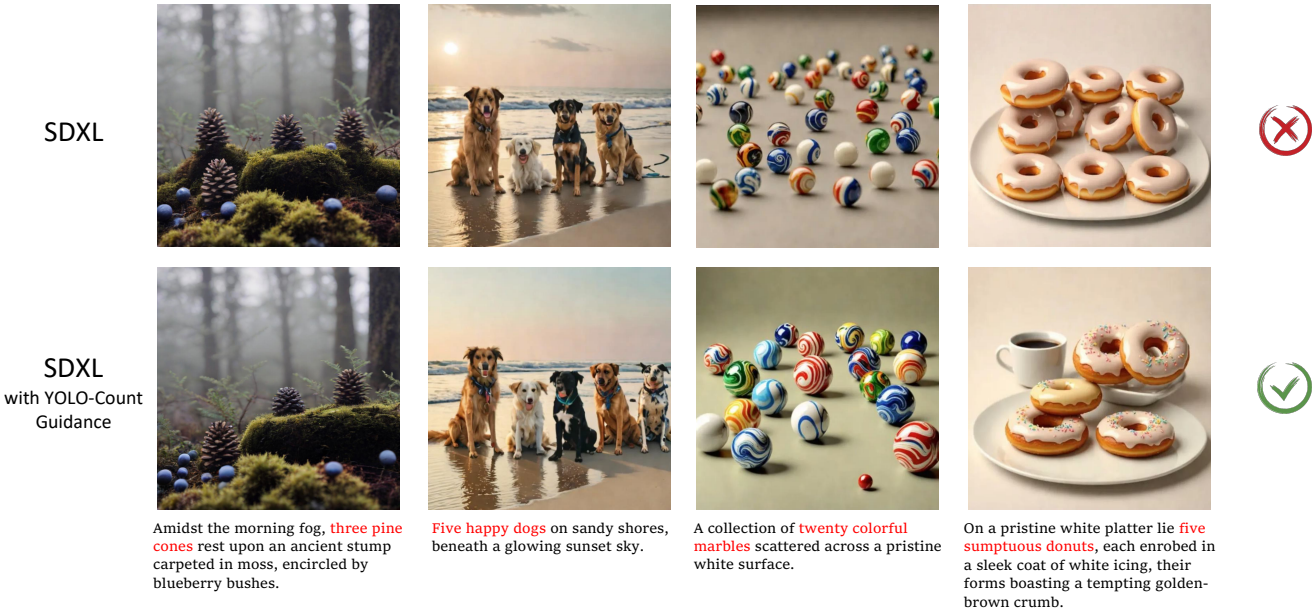

Figure 1. **Demonstrations of YOLO-Count's object quantity controllability** for text-to-image generation. Incorporating YOLO-Count as a differentiable guidance module over a strong baseline (SDXL [41]) substantially improves alignment between text-specified object quantities and the generated images.

Abstract

We propose YOLO-Count, a differentiable open-vocabulary object counting model that tackles both general counting challenges and enables precise quantity control for text-to-image (T2I) generation. A core contribution is the 'cardinality' map, a novel regression target that accounts for variations in object size and spatial distribution. Leveraging representation alignment and a hybrid strong-weak supervision scheme, YOLO-Count bridges the gap between open-vocabulary counting and T2I generation control. Its fully differentiable architecture facilitates gradient-based optimization, enabling accurate object count estimation and fine-grained guidance for generative models. Extensive experiments demonstrate that YOLO-Count achieves state-of-the-art counting accuracy while providing robust and effective quantity control for T2I systems.

[*]Work done during internship at UC San Diego.

1. Introduction

Text-to-image (T2I) generative models have achieved remarkable success in producing high-fidelity images from natural language descriptions. However, ensuring precise alignment with textual specifications, particularly regarding object quantity, remains a significant challenge. While prior research has improved adherence to object layout, attributes, and style through conditional training and guidance mechanisms, accurately controlling the number of objects synthesized within an image remains difficult. Unlike localized attributes, object quantity constitutes a global constraint, requiring models to establish numerical correspondence between language tokens and compositional objects. Consequently, conventional conditional training approaches such as ControlNet [52] are ill-suited for explicit quantity control. Moreover, the stochastic nature of the denoising process in T2I models introduces ambiguity in object differentiation, further complicating count consistency. Re-

cent conditional guidance methods, such as BoxDiff [46] and Ranni [12], address aspects of spatial layout, object attributes, and semantic panel conditioning. However, these methods lack a direct and principled mechanism for precise quantity control, leaving a critical gap in bridging linguistic numeracy and visual synthesis.

In this work, we propose YOLO-Count, an open-vocabulary object counting model built on the YOLO architecture. YOLO-Count is a fully differentiable, regression-based model that demonstrates high accuracy, computational efficiency, and open-vocabulary capabilities. A key contribution is the introduction of the *cardinality map*, a novel representation that encodes object quantity while preserving awareness of object size and spatial location. Unlike traditional density maps, which apply Gaussian kernels at object centers, the cardinality map distributes quantity scores across object instances, improving accuracy and robustness to scale variation. Furthermore, YOLO-Count leverages representation alignment and a hybrid strong-weak supervision strategy, enabling the use of large-scale instance segmentation datasets without reliance on computationally expensive pre-trained visual encoders.

Beyond generic object counting, we are motivated to apply YOLO-Count for precise control of object quantities in text-to-image (T2I) generation. This is achieved by employing YOLO-Count as a differentiable guidance module [5], where gradient signals from the counting model steer the generative process toward numerical consistency. While prior research has predominantly focused on guidance algorithms for attributes and layout, explicit quantity control remains underexplored. We argue that an ideal object counting model for T2I applications should possess four key properties: (1) *full differentiability* w.r.t. the input image; (2) *open-vocabulary* capability for diverse object categories; (3) *cross-scale generalization* to varying object sizes; (4) *computational efficiency* for practical deployment.

Constructing such a model introduces several challenges. First, state-of-the-art counting approaches [2, 35] are often detection-based, producing outputs that preclude gradient propagation. Second, existing counting datasets such as FSC147 [37] or CARPK [18] are limited in scale and category diversity, hindering open-vocabulary generalization. Third, while large-scale vision encoders (*e.g.*, CLIP [36] or GroundingDINO [32, 38]) can alleviate data limitations, they impose significant computational overhead.

To address these issues, we integrate YOLO-Count with textual inversion [13, 50] to achieve precise quantity control in T2I generation. Extensive experiments demonstrate that YOLO-Count achieves state-of-the-art accuracy on counting benchmarks, outperforms density-based and detection-based counting models, and substantially improves object quantity controllability in T2I generation.

Our contributions are summarized as follows:

- We introduce the *cardinality map*, a novel regression target that improves object counting accuracy compared to density maps.
- We develop YOLO-Count, an efficient, open-vocabulary, and *fully differentiable* counting model that achieves state-of-the-art performance and enhances quantity control for T2I generation.
- We propose *hybrid strong-weak supervision* with representation alignment, enabling effective training using large-scale segmentation datasets without reliance on heavy visual encoders.

2. Related Works

2.1. Object Counting Models and Datasets

Object counting models can be broadly classified according to their category scope into fixed-category counting models [14, 40, 44] and open-vocabulary counting models [2, 11, 35]. For controlling object quantities in generative tasks, open-vocabulary counting is essential, as it supports arbitrary object categories without retraining. Based on the type of supervision or guidance, counting models can be further divided into text-guided models [47, 58], visual-exemplar-guided models [20, 37], multimodal-guided models [2], and reference-less models [17, 31, 45]. For T2I integration, a purely text-guided counting model is preferable to ensure compatibility with prompt-driven generation.

From a methodological perspective, counting models are typically divided into detection-based and regression-based approaches. Detection-based models [2, 18, 34] rely on explicit object detection, filtering instances via thresholds and enumerating discrete counts, which inherently produce non-differentiable integer outputs. In contrast, regression-based models [3, 4, 27] predict continuous-valued maps such as density maps [10, 33] that represent pixel-wise contributions to the final count. This direct differentiability makes regression-based models particularly suitable for gradient-based control in generative pipelines.

Finally, training datasets for object counting are categorized into fixed-category datasets [18, 21, 43] and open-vocabulary datasets [1, 37]. Open-vocabulary datasets provide images containing diverse object categories and instance counts, but are expensive to collect and annotate [37]. For example, the widely used FSC147 dataset includes only 3,659 training images, which limits scale and diversity. To address this, recent works [2, 22] incorporate large-scale pre-trained visual backbones (*e.g.*, CLIP [36] and GroundingDINO [32]) and fine-tune them on smaller counting datasets to enhance open-vocabulary generalization.

2.2. Controllable Text-to-Image Generation

Controllable text-to-image (T2I) generation methods can be broadly categorized into two paradigms: training-based

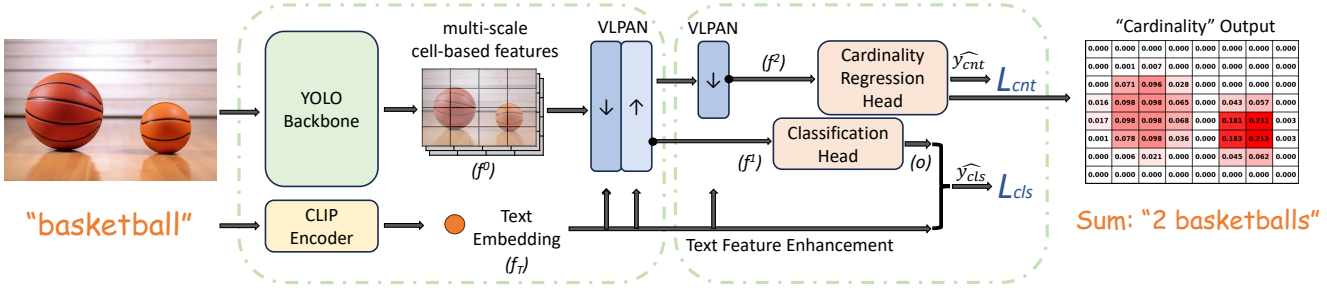

Figure 2. **YOLO-Count Model Overview**. YOLO-Count comprises a YOLO backbone, CLIP text encoder, vision-language path aggregation network (VLPAN), cardinality regression head and classification head. Built upon the YOLO-World [9] architecture, the cardinality head predicts a cardinality map. The final object quantity is obtained by summing over the cardinality map.

methods [19, 52, 54] and guidance-based methods [5, 49, 53]. Training-based approaches, such as ControlNet [52], IP-Adapter [48], and GLIGEN [28], inject conditional inputs directly into the generative model through additional network branches or adapters. While effective, these methods rely on large-scale training datasets annotated with the corresponding conditions. In contrast, guidance-based approaches, including BoxDiff [46], Attend-and-Excite [8], and Separate-and-Enhance [6], control generation by manipulating the diffusion process at inference time, eliminating the need for retraining. Many of these methods exploit the interpretability of cross-attention mechanisms [30] to steer image synthesis. However, cross-attention is primarily effective for distinguishing object categories rather than differentiating multiple instances of the same category. As a result, existing controllable T2I techniques excel at localized attribute binding [15, 55] and layout control [56, 57], but struggle with enforcing global constraints such as precise object quantity.

2.3. Object Quantity Control for T2I Models

Research on explicit object quantity control in text-to-image (T2I) models remains limited. [25] pioneers the use of universal diffusion guidance for quantity control, representing the first attempt to directly address this challenge. [7] introduces an attention-based representation for counting objects, but their approach is constrained to controlling small quantities (ranging from 1 to 10). More recently, prompt-tuning approaches [42, 50] have been proposed to incorporate numerical cues into the text embedding space, enabling limited quantity control without modifying the underlying diffusion model. However, these methods still struggle with accurate control over larger counts.

3. Methods

3.1. Model Overview

Our proposed YOLO-Count builds upon the YOLO-World architecture [9] and consists of three primary components: (1) a vision backbone, (2) a vision-language path aggregation network (VLPAN), and (3) prediction heads. Fig. 2 illustrates the overall pipeline and highlights our key architectural modifications.

Vision Backbone. The vision backbone in YOLO-Count follows the design of YOLOv8l [23] and YOLO-World-L [9]. It comprises five stages of convolutional modules (ConvModules) and cross-stage partial layers (CSPLayers). Given an input image $I \in \mathbb{R}^{640\times640\times3}$, the backbone extracts multiscale visual features at three resolutions:

$$f^0 = [f_{80\times80}, f_{40\times40}, f_{20\times20}] = \text{VisualBackbone}(I)$$

Vision-Language Path Aggregation Network (VLPAN). The VLPAN is designed to fuse visual features with textual semantics and aggregate information across scales. Inheriting from YOLO-World, it employs both top-down and bottom-up pathways, but with key enhancements: (1) T-CSPLayers: standard CSPLayers are replaced by T-CSPLayers, which integrate sigmoid attention blocks to modulate visual features based on precomputed CLIP text embeddings [36]. (2) Extended Top-Down Fusion: to better preserve fine-grained spatial details, an additional top-down pathway is introduced following the initial bidirectional aggregation, maximizing high-resolution feature utilization, which is critical for accurate counting regression. The enhanced VLPAN is formulated as:

$$[f^1, f^2] = \text{VLPAN}(f^0, f_\text{T})$$

where f_T denotes the CLIP text embedding of the category, f^1 and f^2 represent multimodal features for classification and counting regression, respectively.

Prediction Heads. Following the VLPAN, several ConvModules are applied to text-aware visual features to aggregate multi-scale signals into a unified 80×80 resolution. The prediction stage then produces two parallel outputs: (1) a cardinality regression head, which predicts a

dense cardinality map for differentiable counting, and (2) a classification head, trained with contrastive supervision to ensure robust open-vocabulary capability. These two outputs jointly enable YOLO-Count to provide accurate, differentiable count estimates while maintaining strong category generalization, as shown on the right side of Fig. 2.

$$\begin{cases} o_{\text{cls}} = \text{ClassificationHead}(f^1) \\ \hat{y}_{\text{cnt}} = \text{CountingHead}(f^2) \end{cases}$$

3.2. Cardinality Map Regression

We introduce the concept of the *cardinality map*, a novel regression target designed to address the inherent ambiguities of density-based counting.

Density-based counting models formulate the counting loss as:

$$\mathcal{L}_{\text{cnt}} = |\hat{y}_{\text{cnt}} - y_{\text{den}}| \tag{1}$$

where y_{den} is the density map. For an image containing Q objects, y_{den} is constructed by placing Q Gaussian kernels centered at object locations, with their sum equal to Q. While regressing to y_{den} enables quantity prediction, this representation suffers from two key ambiguities. First, the Gaussian kernel's center can be placed anywhere within an object's extent. Second, the kernel radius is arbitrarily chosen and lacks physical meaning. These issues degrade model accuracy for objects with diverse sizes and shapes, as density maps fail to provide a consistent, unambiguous representation.

To overcome these limitations, we replace y_{den} with a *cardinality map* y_{car}, defined using object masks. Given a binary mask M_i for the i-th object instance, with area $N_i = |M_i|$, we uniformly distribute a value of 1 across all pixels within the object and sum contributions across all K objects:

$$y_{\text{pixel cardinality}} = \sum_{i=1}^{K} \frac{1}{N_i} M_i \tag{2}$$

We then downsample this pixel-level cardinality map to a grid-based representation by summing within each grid cell:

$$y_{\text{car}}(u, v) = \sum_{(i,j) \in \Omega_{u,v}} y_{\text{pixel cardinality}}(i, j) \tag{3}$$

where $\Omega_{u,v}$ is the set of pixel coordinates within grid cell (u, v). By construction, the total sum of the cardinality map equals the true object count:

$$\sum_{u,v} y_{\text{car}}(u, v) = Q \tag{4}$$

Unlike density maps, which concentrate mass at object centers and often ignore large portions of extended objects, the cardinality map uniformly covers the entire spatial extent of each object. This yields a unique, unambiguous representation that is robust to variations in object size and shape, making it better suited for differentiable regression-based counting.

3.3. Representation Alignment

Here we describe our method using the single-category counting scenario for simplicity, where the model is designed to count instances of a specific category specified by the user. To achieve this, we adapt the contrastive learning framework into a binary classification task, where each pixel is classified as either belonging to the target category or not. This additional branch aligns the visual and textual representations during training, ensuring the model effectively localizes instances of the specified category. Specifically, the classification loss is formulated as:

$$\mathcal{L}_{\text{cls}} = \text{BCELoss}(\hat{y}_{\text{cls}}, y_{\text{cls}}) \tag{5}$$

where $y_{\text{cls}}(i, j) \in \{0, 1\}$ is the binary ground truth label indicating whether pixel (i, j) belongs to the target category, and $\hat{y}_{\text{cls}}(i, j) \in [0, 1]$ is the predicted probability obtained by projecting the visual feature o_{cls} and the text embedding f_{T} into a shared multimodal space and applying a sigmoid activation to their inner product, similar to SigLIP [51].

3.4. Hybrid Strong-Weak Training

Training an object-counting model typically necessitates a specialized counting dataset, where each image contains multiple instances of a single category. Regression-based counting models rely on corresponding density maps, which are costly to produce and task-specific. To overcome this data limitation, we propose a hybrid strong-weak training technique that enables cardinality regression using both instance segmentation datasets and counting datasets. This approach comprises two stages: strong-supervision pretraining and weak-supervision finetuning.

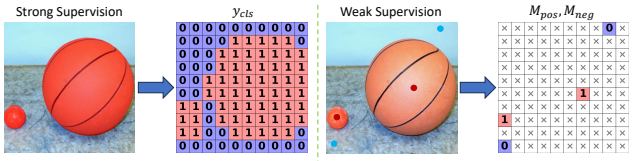

Figure 3. **Illustration of hybrid strong-weak supervision.** Instance segmentation masks provide dense, per-grid category labels, while counting annotations offer sparse category labels limited to annotated points. By combining these two forms of supervision, we enable effective training of the object counting model using both precise (strong) and sparse (weak) annotations.

3.4.1. Strong-Supervision Pretraining

We first pretrain the model on instance segmentation datasets where precise per-instance masks are available. We

construct cardinality maps y_{car} (as defined in Eq. (3)) and binary classification masks $y_{\text{cls}} \in \{0,1\}^{H\times W}$. The pretraining objectives are as follows:

$$\mathcal{L}_{\text{cnt}}^{\text{strong}} = |\hat{y}_{\text{cnt}} - y_{\text{car}}| \tag{6}$$

$$\mathcal{L}_{\text{total}}^{\text{strong}} = \alpha_1 \mathcal{L}_{\text{cnt}}^{\text{strong}} + \beta_1 \mathcal{L}_{\text{cls}}^{\text{strong}} \tag{7}$$

where α_1 and β_1 are weighting coefficients. $\mathcal{L}_{\text{cls}}^{\text{strong}}$ is the same $\mathcal{L}_{\text{cls}}$ defined in Eq. (5). This stage provides precise pixel-level supervision for both cardinality regression and category-specific classification, establishing a strong model initialization.

3.4.2. Weak-Supervision Finetuning

To better adapt YOLO-Count to dense counting scenarios, we perform weak-supervision finetuning on counting datasets that provide sparse point-level annotations. Each image is annotated with instance points $\mathcal{P} = \{(x_i, y_i)\}_{i=1}^{K}$, where K is both the total count of objects in the image and the number of annotated points. Weak supervision comprises the following two components:

(1) Sparse Classification Labels: Positive labels are derived from annotated points, forming $M_{\text{pos}} \in \{0,1\}^{H\times W}$ where $M_{\text{pos}}(i,j) = 1$ if $(i,j) \in \mathcal{P}$. Negative labels M_{neg} are sampled from background regions.

(2) Total Count Consistency: The predicted total count must match the ground truth count K.

The weak supervision losses are then formulated as follows:

$$\mathcal{L}_{\text{cls}}^{\text{weak}} = -\frac{1}{|\Omega|} \sum_{p\in\Omega} \big[M_{\text{pos}}(p) \log \hat{y}_{\text{cls}}(p) + M_{\text{neg}}(p) \log\left(1 - \hat{y}_{\text{cls}}(p)\right) \big] \tag{8}$$

$$\mathcal{L}_{\text{cnt}}^{\text{weak}} = \left| \left(\sum_p \hat{y}_{\text{cnt}}(p) \right) - K \right| \tag{9}$$

$$\mathcal{L}_{\text{total}}^{\text{weak}} = \alpha_2 \mathcal{L}_{\text{cnt}}^{\text{weak}} + \beta_2 \mathcal{L}_{\text{cls}}^{\text{weak}}, \tag{10}$$

where $\Omega = \{(i,j) \,|\, M_{\text{pos}}(i,j) = 1 \vee M_{\text{neg}}(i,j) = 1\}$ denotes the annotated pixel locations (see Fig. 3).

This hybrid training scheme effectively leverages large-scale instance segmentation datasets for pretraining and adapts to limited counting datasets during finetuning, enabling robust and data-efficient model training.

3.5. Counting-Controlled Generation

Following [50], we adopt a textual inversion-based approach for counting-controlled generation. The text-to-image (T2I) model first synthesizes an initial image from a text prompt, which may not accurately reflect the desired object quantity. We then feed the generated image and its target category into the YOLO-Count model to estimate the predicted quantity and compute a guidance loss based on the deviation from the required count Q_{req}:

$$\mathcal{L}_{\text{guide}} = \left| \left(\sum_p \hat{y}_{\text{cnt}}(p) \right) - Q_{\text{req}} \right| \tag{11}$$

We iteratively update a learnable counting token inserted into the text sequence via the gradient of $\mathcal{L}_{\text{guide}}$. This process continues until convergence, effectively steering the T2I model toward images with the desired object quantity.

4. Experiments

4.1. Setups

4.1.1. Training Datasets

FSC147 [37] is an object counting dataset comprising 6,135 images across 89 training classes, 29 validation classes, and 29 test classes, with no overlap between the splits. Following [2], we applied corrections to several images containing erroneous annotations to ensure label consistency.

LVIS v1.0 [16] is a large-scale, long-tailed dataset containing 1,203 object categories, sharing images with MSCOCO [29]. Originally designed for instance segmentation with a large vocabulary, we employ its validation set to evaluate counting accuracy across diverse and open-vocabulary categories.

4.1.2. Evaluation Metrics

Following prior work [18, 21, 37], we employ mean absolute error (MAE) and root mean square error (RMSE) to evaluate counting performance. MAE measures the average absolute deviation between predicted and ground-truth counts, while RMSE penalizes larger errors more heavily, capturing both accuracy and precision.

4.1.3. Benchmarks of Object Counting

In addition to FSC147 and LVIS, we introduce two new benchmarks to evaluate open-vocabulary object counting accuracy.

OpenImages V7 [26] and Objects365 [39] are large-scale open-vocabulary object detection datasets, containing 9 million images across 600 categories and 2 million images across 365 categories, respectively. Based on these datasets, we construct two novel counting benchmarks: **OpenImg7-New** and **Obj365-New**, specifically designed to evaluate generalization to unseen categories.

To ensure category novelty, we first compute CLIP [36] embeddings for all text labels in both datasets. We then filter out categories whose maximum cosine similarity with any LVIS label embedding exceeds 0.7. This procedure yields 47 novel categories from OpenImages V7 and 51 from Objects365. Finally, we retain only images containing these filtered categories, resulting in benchmarks with 14,699 images (OpenImg7-New) and 22,724 images (Obj365-New).

Model	FSC-Test [37] *(Seen Categories)*		FSC-Val [37] *(Seen Categories)*		LVIS [16] *(Seen for Ours)*		OImg7-New [26] *(Unseen Categories)*		Obj365-New [39] *(Unseen Categories)*	
	MAE↓	RMSE ↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
Non-Differentiable										
CountGD [2]	12.98	98.35	12.14	47.51	4.84	12.45	6.09	29.92	3.53	10.61
DAVE [35]	14.90	103.42	15.48	52.57	5.29	11.48	5.31	14.24	4.89	13.22
Differentiable										
CLIP-Count [22]	17.78	106.62	18.79	61.18	10.81	22.61	14.01	30.16	15.48	30.28
VLCounter [24]	17.05	106.16	18.06	65.13	8.94	23.04	15.32	34.41	18.08	33.02
CounTX [1]	15.88	106.29	17.10	65.61	13.70	40.94	17.94	48.73	18.76	44.80
YOLO-Count **(Ours)**	**14.80**	**96.14**	**15.43**	**58.36**	**1.65**	**6.08**	**3.72**	**11.96**	**3.28**	**9.15**

Table 1. **Comparison of counting accuracy** with existing text-guided object counting models.

These benchmarks provide a challenging evaluation protocol for assessing the open-vocabulary generalization ability of object counting models beyond the categories seen during FSC147 and LVIS training.

4.1.4. Benchmarks of Controllable Generation

To evaluate the precision of quantity-controlled text-to-image (T2I) generation, we construct two benchmarks: **LargeGen** and **LargeGen-New**. LargeGen is derived from the FSC147 dataset by selecting the 10 text categories with the highest average object counts per image, providing a benchmark for evaluating generation performance on seen categories. LargeGen-New is built from Obj365-New and OpenImg7-New, thereby assessing generation accuracy on novel categories. These benchmarks enable a systematic evaluation of quantity-guided generation, measuring both seen-category and novel-category performance when using object counting models to steer T2I generation.

4.1.5. Training YOLO-Count Model

For YOLO-Count, we first perform a 250-epoch strong-supervision pretraining on the LVIS dataset. The vision backbone is initialized with YOLOv8l weights, while all other modules are randomly initialized. We optimize a composite loss with coefficients $\alpha_1 = 1.0$ for cardinality regression and $\beta_1 = 0.1$ for category classification. The CLIP text encoder remains frozen during training, while all other parameters are updated with different learning rates: 5×10^{-9} for the backbone (to preserve pre-trained visual representations) and 1×10^{-5} for newly initialized modules. This learning rate strategy stabilizes training and facilitates adaptation of the YOLO backbone to the counting task.

Subsequently, we perform weak-supervised finetuning on FSC147, leveraging the dataset-provided positive point annotations and manually annotating negative labels. For negative labels, we dot approximately 10 background points per image in the FSC147 training set. This annotation process is highly efficient, facilitated through a labeling interface, and requires only about 5 seconds per image on average. During finetuning, we retain a proportion γ of LVIS data within each training batch to provide strong supervision and preserve the open-vocabulary capability of YOLO-Count. The model is trained for up to 500 epochs with early stopping based on the mean absolute error (MAE) on the FSC147 validation set.

4.1.6. Counting Control with Token Optimization

We integrate YOLO-Count with SDXL-Turbo [41] using a differentiable token optimization strategy for quantity-controlled generation. The pipeline employs single-step inference to balance generation quality and computational efficiency. For each generation task, we iteratively optimize a count token embedding for up to 150 steps using a learning rate of 5×10^{-3}, with early stopping applied if the guidance loss (Eq. (11)) plateaus for 20 consecutive steps. This process allows direct gradient-based feedback from YOLO-Count to refine the token embedding, effectively steering the T2I model toward producing images with the desired object quantity.

4.2. Results of Object Counting

We evaluate counting accuracy on the FSC147 validation and test sets, a widely used benchmark for object-counting models. For FSC147, we set $\alpha_1 = 0$, $\beta_1 = 1$, $\alpha_2 = 1$, $\beta_2 = 0.1$, and $\gamma = 0.05$ during finetuning. Following prior work such as CountGD [2] and CLIP-Count [22], we adopt an automatic clipping and aggregation strategy to handle

Figure 4. Counting results of YOLO-Count.

images containing large object counts. As shown in Tab. 1, YOLO-Count achieves state-of-the-art performance among regression-based counting models. Our MAE and RMSE scores on FSC147 are competitive and closely approach those of CountGD, the current leading non-differentiable counting model. Visual examples in Fig. 4 further demonstrate that YOLO-Count produces accurate object counts.

We further evaluate the open-vocabulary capability of YOLO-Count by testing its counting accuracy on unseen categories. For LVIS, OpenImg7-New, and Objects365-New, we set $\alpha_1 = \beta_1 = \alpha_2 = \beta_2 = 1$ and $\gamma = 0.5$ for fine-tuning. The results, shown in Tab. 1, demonstrate that exposure to more diverse training data significantly improves YOLO-Count's ability to generalize to novel categories. Notably, despite having the fewest parameters among the compared counting models, YOLO-Count achieves superior accuracy in open-vocabulary settings.

In summary, Figs. 1 and 4 to 6 illustrate results obtained using the FSC147 checkpoint settings of Tab. 1. Figs. 7 and 8 present results based on the checkpoint settings for LVIS, OpenImg7, and Objects365 of Tab. 1.

4.3. Results of T2I Quantity Control

We evaluate the effectiveness of YOLO-Count in guiding T2I models for accurate quantity control. We compare against two series of baselines: (1) extension of previous guidance-based methods [7, 25] to large quantities, and (2) the token optimization framework of [50] guided by alternative counting models, including CLIP-Count and CountGD.

For each category in LargeGen and LargeGen-New, we generate 10 images per target quantity, with counts set to 25, 50, 75, and 100. Notably, since CountGD is non-differentiable, we cannot apply the gradient-based loss in Eq. (11). Instead, we employ a surrogate cross-entropy objective that encourages a high probability for the target count Q_{req} while suppressing probabilities for other counts.

We manually evaluate the generated images by counting the objects and comparing them against the target Q_{req}. As shown in Fig. 5, YOLO-Count substantially improves generation accuracy compared to all baselines.

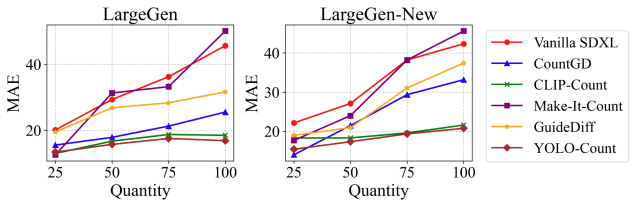

Figure 5. **Quantitive results for T2I quantity control.** Our method significantly reduces the discrepancy between the requested and generated object quantities compared to previous approaches, both on seen and unseen categories.

Furthermore, Fig. 6 illustrates qualitative differences when generating scenes with large object counts. Under CountGD's surrogate loss $\mathcal{L}^{\text{det}}_{\text{guide}}$, the T2I model fails to adjust object counts and suffers from degraded visual quality. Similarly, density-based counting guidance results in large deviations from Q_{req}, likely due to the domain gap between real counting datasets and synthetic T2I images. In contrast, YOLO-Count provides precise, differentiable guidance signals, enabling accurate quantity control while generalizing effectively to novel categories. Additional qualitative examples are provided in the *Supplementary Material*.

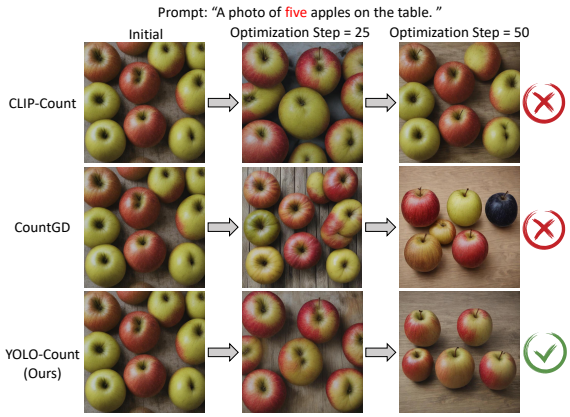

Figure 6. **Comparison of counting-controlled generation** by different models.

4.4. Ablation Study

We conduct an ablation study to evaluate the contribution of each key component in YOLO-Count. Specifically, we examine the impact of our cardinality map regression, representation alignment, hybrid strong-weak training strategy, and architectural modifications on the performance of regression-based object counting models.

Model	FSC-Test [37]		FSC-Val [37]	
	MAE↓	RMSE↓	MAE↓	RMSE↓
Baseline				
YOLO-Count	14.80	96.14	15.43	58.36
...without				
Pretraining	18.42	111.45	19.50	88.64
Weak-Supervision	43.91	150.40	43.86	124.33
Cardinality Map	16.71	107.24	17.87	76.42
Alignment	17.01	110.41	17.57	85.54
Additional VLPAN	16.54	106.32	16.89	84.40

Table 2. **Ablation Studies.** w/o Pretraining: no LVIS data; w/o Weak-Supervision: no FSC147 data; w/o Cardinality Map: training on density maps directly; w/o Alignment: no classification branch; w/o Additional VLPAN: both YOLO-Count heads receive same VLPAN feature output.

We begin by evaluating the contribution of each stage in

the training pipeline. Specifically, we analyze counting accuracy under two conditions: (1) without strong-supervised pretraining and (2) without weak-supervised finetuning. In the first scenario (w/o pretraining), YOLO-Count is trained directly on FSC147 using density map supervision. In the second scenario (w/o finetuning), YOLO-Count is trained solely on LVIS with strong labels, without subsequent finetuning. As shown in Tab. 2, the absence of finetuning leads to significantly higher MAE and RMSE, highlighting the domain gap between LVIS and FSC147 in terms of object count distributions. Importantly, combining strong-supervised pretraining on LVIS with weak-supervised finetuning on FSC147 achieves substantial improvements over either stage alone, validating our hybrid training strategy.

Next, we examine the impact of removing the cardinality regression component. In this variant (w/o cardinality), we pretrain on LVIS as in the default setting, but replace cardinality map regression with density map regression during FSC147 training. As observed in Tab. 2, the inclusion of cardinality regression yields consistently lower MAE and RMSE compared to density-based regression. Furthermore, visual results in Fig. 7 reveal that density-based regression often suffers from overlapping kernels within a single instance and exhibits a bias toward over-counting in images containing larger-sized objects. This demonstrates that cardinality regression mitigates ambiguities inherent in density maps and improves robustness across varying object scales.

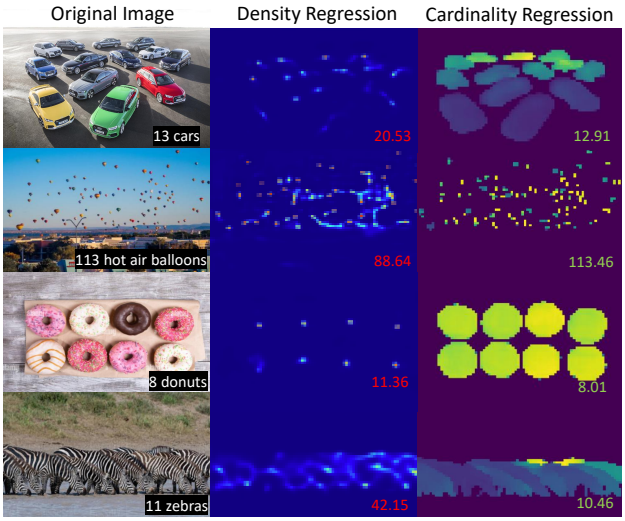

Figure 7. **Comparison of density *vs*. cardinality regression**.

4.5. Analysis on Size Bias of Counting Models

We design an experiment to investigate the size bias exhibited by density-based regression methods when handling objects of varying scales. Specifically, we select images from the FSC147 validation and test sets containing no more than 30 objects. Each image is progressively downscaled with a scaling ratio ranging from 1.0 (original size) to 4.0, and then padded to restore the original image dimensions. While this process reduces the object sizes, the ground-truth counts remain unchanged.

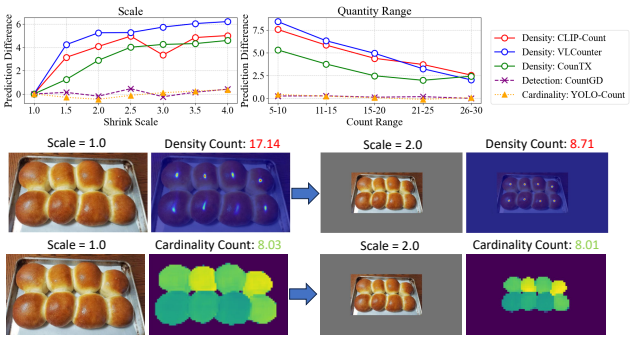

Figure 8. **Comparison of counting bias** for detection, density-based regression and cardinality-based regression models in different image downscaling scales and number of objects.

The processed images are then fed into several object-counting models: three density-based models (VLCounter [24], CLIP-Count [22], and CounTX [1]), one detection-based model (CountGD [31]), and our proposed YOLO-Count model. For each model, we record the predicted counts and analyze their differences relative to predictions on the original (unscaled) images, providing a measure of counting stability under size variation.

As illustrated in Fig. 8, density-based regression models consistently over-count as object sizes increase, with this bias becoming more pronounced for images containing larger objects. In contrast, YOLO-Count exhibits stability similar to detection-based models, maintaining accurate counts across different object scales. This result underscores the advantage of training regression-based models on cardinality maps, which eliminates kernel-based ambiguities and improves robustness to variations in object size.

5. Conclusion

In this paper, we introduced YOLO-Count, a novel open-vocabulary, regression-based object counting model that substantially improves object quantity control in text-to-image generation. By integrating cardinality regression, hybrid strong-weak supervision, and representation alignment, YOLO-Count achieves state-of-the-art counting accuracy, computational efficiency, and robust open-vocabulary generalization. Extensive experiments and ablation studies validate its effectiveness in overcoming the limitations of prior approaches, particularly in handling large object quantities and novel categories. Beyond advancing object counting, YOLO-Count provides a practical and differentiable mechanism for enhancing the controllability of text-to-image models, thereby enabling more precise and reliable multimodal generation and perception.

Acknowledgment This work is supported by NSF Award IIS-2127544 and NSF Award IIS-2433768. We thank Yuheon Joh for insightful discussions and valuable feedback.

References

[1] Niki Amini-Naieni, Kiana Amini-Naieni, Tengda Han, and Andrew Zisserman. Open-world text-specifed object counting. In *BMVC*, page 510, 2023. 2, 6, 8

[2] Niki Amini-Naieni, Tengda Han, and Andrew Zisserman. Countgd: Multi-modal open-world counting. In *Advances in Neural Information Processing Systems*, pages 48810–48837. Curran Associates, Inc., 2024. 2, 5, 6

[3] Carlos Arteta, Victor Lempitsky, J Alison Noble, and Andrew Zisserman. Interactive object counting. In *European Conference on Computer Vision*, pages 504–518. Springer, 2014. 2

[4] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. In *European Conference on Computer Vision*, pages 483–498. Springer, 2016. 2

[5] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 843–852, 2023. 2, 3

[6] Zhipeng Bao, Yijun Li, Krishna Kumar Singh, Yu-Xiong Wang, and Martial Hebert. Separate-and-enhance: Compositional finetuning for text-to-image diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 3

[7] Lital Binyamin, Yoad Tewel, Hilit Segev, Eran Hirsch, Royi Rassin, and Gal Chechik. Make it count: Text-to-image generation with an accurate number of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13242–13251, 2025. 3, 7

[8] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Trans. Graph.*, 42(4), 2023. 3

[9] 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 (CVPR)*, pages 16901–16911, 2024. 3

[10] Siu-Yeung Cho, Tommy WS Chow, and Chi-Tat Leung. A neural-based crowd estimation by hybrid global learning algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(4):535–541, 1999. 2

[11] Siyang Dai, Jun Liu, and Ngai-Man Cheung. Referring expression counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16985–16995, 2024. 2

[12] Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. Ranni: Taming text-to-image diffusion for accurate instruction following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4744–4753, 2024. 2

[13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[14] Mingyue Guo, Li Yuan, Zhaoyi Yan, Binghui Chen, Yaowei Wang, and Qixiang Ye. Regressor-segmenter mutual prompt learning for crowd counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28380–28389, 2024. 2

[15] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9380–9389, 2024. 3

[16] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5, 6

[17] Michael Hobley and Victor Prisacariu. Learning to count anything: Reference-less class-agnostic counting with weak supervision. *arXiv preprint arXiv:2205.10203*, 2022. 2

[18] 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 (ICCV)*, 2017. 2, 5

[19] Minghui Hu, Jianbin Zheng, Daqing Liu, Chuanxia Zheng, Chaoyue Wang, Dacheng Tao, and Tat-Jen Cham. Cocktail: Mixing multi-modality control for text-conditional image generation. In *Advances in Neural Information Processing Systems*, pages 32424–32444. Curran Associates, Inc., 2023. 3

[20] Zhizhong Huang, Mingliang Dai, Yi Zhang, Junping Zhang, and Hongming Shan. Point segment and count: A generalized framework for object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17067–17076, 2024. 2

[21] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 5

[22] 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*, pages 4535–4545, 2023. 2, 6, 8

[23] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 3

[24] 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*, pages 2714–2722, 2024. 6, 8

[25] Wonjun Kang, Kevin Galim, Hyung Il Koo, and Nam Ik Cho. Counting guidance for high fidelity text-to-image synthesis.

In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 899–908, 2025. 3, 7

[26] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 5, 6

[27] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2010. 2

[28] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22511–22521, 2023. 3

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5

[30] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7817–7826, 2024. 3

[31] Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. Countr: Transformer-based generalised visual counting. *arXiv preprint arXiv:2208.13721*, 2022. 2, 8

[32] 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*, pages 38–55. Springer, 2024. 2

[33] Aparecido Nilceu Marana, Sergio A Velastín, LF Costa, and RA Lotufo. Estimation of crowd density using image processing. In *IEE Colloquium on Image Processing for Security Applications (Digest No: 1997/074)*, pages 11–1. IET, 1997. 2

[34] Thanh Nguyen, Chau Pham, Khoi Nguyen, and Minh Hoai. Few-shot object counting and detection. In *European Conference on Computer Vision*, pages 348–365. Springer, 2022. 2

[35] Jer Pelhan, Alan Lukeži?, Vitjan Zavrtanik, and Matej Kristan. Dave - a detect-and-verify paradigm for low-shot counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23293–23302, 2024. 2, 6

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5

[37] 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 (CVPR)*, pages 3394–3403, 2021. 2, 5, 6, 7

[38] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the" edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024. 2

[39] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 5, 6

[40] Zenglin Shi, Pascal Mettes, and Cees GM Snoek. Focus for free in density-based counting. *International Journal of Computer Vision*, 132(7):2600–2617, 2024. 2

[41] Stability AI. Sdxl-turbo: Text-to-image generative model, 2024. 1, 6

[42] Wenfang Sun, Yingjun Du, Gaowen Liu, and Cees GM Snoek. Quota: Quantifying objects with text-to-image models for any domain. *arXiv preprint arXiv:2411.19534*, 2024. 3

[43] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[44] Yi Wang, Qiongyang Hu, and Lap-Pui Chau. Weakly-supervised crowd counting with token attention and fusion: A simple and effective baseline. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13456–13460. IEEE, 2024. 2

[45] Yuzhe Wu, Yipeng Xu, Tianyu Xu, Jialu Zhang, Jianfeng Ren, and Xudong Jiang. Gca-sun: A gated context-aware swin-unet for exemplar-free counting. *arXiv preprint arXiv:2409.12249*, 2024. 2

[46] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7452–7461, 2023. 2, 3

[47] 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 (CVPR)*, pages 15548–15557, 2023. 2

[48] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[49] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23174–23184, 2023. 3

[50] Oz Zafar, Lior Wolf, and Idan Schwartz. Iterative object count optimization for text-to-image diffusion models. *arXiv preprint arXiv:2408.11721*, 2024. 2, 3, 5, 7

[51] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023. 4

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 1, 3

[53] Peiang Zhao, Han Li, Ruiyang Jin, and S Kevin Zhou. Loco: Locally constrained training-free layout-to-image synthesis. *arXiv preprint arXiv:2311.12342*, 2023. 3

[54] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, pages 11127–11150. Curran Associates, Inc., 2023. 3

[55] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, BIN FU, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. In *Advances in Neural Information Processing Systems*, pages 73969–73982. Curran Associates, Inc., 2023. 3

[56] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22490–22499, 2023. 3

[57] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6818–6828, 2024. 3

[58] 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*, pages 368–385. Springer, 2024. 2