ABACUS: Adapting Unified Foundation Model for Bridging Image Count Understanding and Generation

ANINDYA MONDAL*, University of Surrey, United Kingdom
SAURADIP NAG*, Simon Fraser University, Canada
ANJAN DUTTA, University of Surrey, United Kingdom

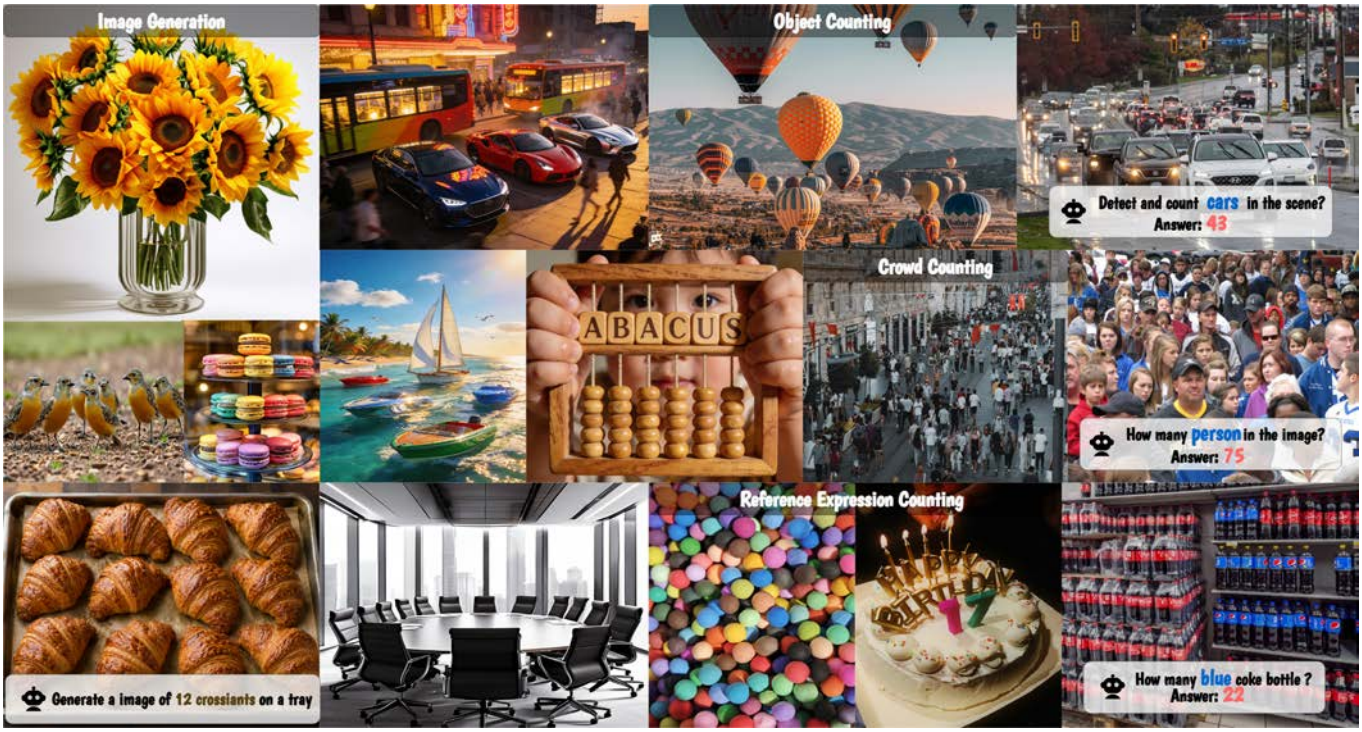

Fig. 1. ABACUS overview. A single unified model performs count-aware image generation (left), object counting, crowd counting, and referring-expression counting (right) using text-only prompts with no benchmark-specific training.

We present ABACUS, a unified vision-language model that jointly addresses object counting, crowd counting, referring-expression counting, and count-faithful image generation within a single 3B-parameter model. ABACUS introduces three contributions: density-aware adaptive zooming paired with an objectness map from multi-head self-attention decomposition to spatially ground count predictions; a boundary-aware count policy trained via GRPO with nested local, boundary, and global rewards to eliminate over- and undercounting at crop boundaries; and a cycle-consistent GRPO strategy in which the frozen understanding branch scores generated candidates on count-deviation and aesthetic quality, closing the understanding–generation synergy gap without any external critic or annotation. ABACUS achieves state-of-the-art results across seven benchmarks spanning object counting (FSC-147, CARPK), crowd counting (ShanghaiTech A/B), referring-expression counting (REC-8K), count-faithful generation (CoCo-Count, T2I-CompBench, GenEval), and count reasoning (CountQA), surpassing both task-specific specialists and larger generalist models. Project page is at https://mondalanindya.github.io/pages/ABACUS/.

ACM Reference Format:
Anindya Mondal*, Sauradip Nag*, and Anjan Dutta. 2026. ABACUS: Adapting Unified Foundation Model for Bridging Image Count Understanding and Generation. *ACM Trans. Graph.* 1, 1 (September 2026), 18 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Authors have equal contributions.

Authors' Contact Information: Anindya Mondal*, University of Surrey, United Kingdom, a.mondal@surrey.ac.uk; Sauradip Nag*, Simon Fraser University, Canada, s.nag@sfu.ca; Anjan Dutta, University of Surrey, United Kingdom, anjan.dutta@surrey.ac.uk.

ACM 1557-7368/2026/9-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Object counting [Amini-Naieni et al. 2023a; Liu et al. 2022; Ranjan et al. 2021] is the task of estimating the number of target instances in an image, conventionally addressed by regressing a density map whose spatial integral gives the total count. Count-conditioned

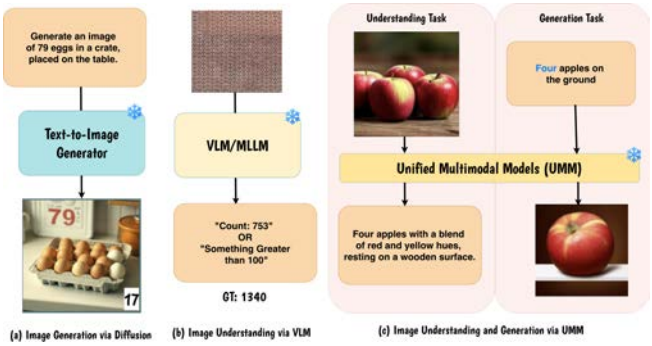

Fig. 2. Issues in Count Generation and Understanding. (a) Text-to-image diffusion models lack any mechanism to verify output cardinality. (b) VLMs and MLLMs default to coarse magnitude estimates (*e.g.*, "Greater than 100") on dense scenes. (c) Existing UMMs support both tasks from a single models yet exhibit a *synergy gap*: the same model that correctly counts 4 apples cannot generate exactly 4.

image generation [Binyamin et al. 2025; Kang et al. 2025] is the complementary task of synthesising a scene containing precisely the number of instances specified in a text prompt. Despite remarkable progress in both directions, these tasks have been pursued in isolation — object counting [Amini-Naieni et al. 2023b; Amini-Naieni and Zisserman 2025; Ranjan et al. 2021], crowd counting [Liang et al. 2023; Shu et al. 2022], referring-expression counting [Dai et al. 2024a], and count-conditioned generation [Binyamin et al. 2025; Kang et al. 2025; Mondal et al. 2025] each rely on task-specific architectures, loss formulations, and training pipelines that do not generalise across regimes. Attempts to unify these under broader vision-language models [Amini-Naieni and Zisserman 2025; Zhang et al. 2026] have shown limited success: such models generalise across object categories but collapse on fine-grained, instance-level instructions due to a lack of spatial grounding during training. At the heart of this difficulty lies *objectness* the coherent representation of a distinct entity even amid visually similar neighbours [Alexe et al. 2012; Kuo et al. 2015] a capacity long studied in cognitive psychology [Spelke 1990] and still poorly captured by current models. On the generation side, state-of-the-art diffusion models [Binyamin et al. 2025; Dahary et al. 2024] routinely produce incorrect cardinalities (see Fig 2(a)) or spatially incoherent arrangements, underscoring that generating a precise number of objects requires reasoning about global spatial relations that purely generative objectives do not enforce. These limitations motivate a unified model capable of jointly performing count understanding and count-faithful generation from a shared representation, without task-specific supervision.

Recent unified multimodal models [Chen et al. 2025b; Deng et al. 2025; Tang et al. 2025; Xie et al. 2024] have demonstrated that visual understanding and generation can be handled within a shared parameter space, yet closing the gap between the two capabilities remains non-trivial. Existing approaches rely on carefully engineered training recipes — including data and loss balancing, multi-stage optimisation, and hybrid MLLM-diffusion pipelines — to prevent gains in one task from degrading the other. Even under such regimes, a persistent *synergy gap* remains: as illustrated in fig. 2(c), the same unified checkpoint that correctly counts four apples in an image fails to generate exactly four apples from a text prompt. This asymmetry is compounded by a more fundamental deficiency, off-the-shelf MLLMs struggle to parse dense visual scenes [Qharabagh et al. 2024] (Fig 2(a)), so the understanding branch of native UMMs cannot provide reliable count supervision for the generation branch. Consequently, neither task benefits from the complementary signal the other could in principle offer.

To bridge this gap, we introduce **ABACUS**, a unified vision-language model built upon [Tang et al. 2025] that jointly addresses object counting, crowd counting, referring-expression counting, and count-faithful image generation within a single unified model, generalising to real-world scenarios in a zero-shot manner. To overcome the well-documented failure of MLLMs on dense scenes [Qharabagh et al. 2024; Zhang et al. 2026], we propose density-aware adaptive zooming, which recursively partitions dense images into manageable sub-regions so that the MLLM operates over locally sparse fields where per-instance discrimination is more reliable for counting. Complementing this, we derive an objectness map from multi-head self-attention decomposition [Vaswani et al. 2017] of the MLLM attention layers that spatially grounds count predictions in genuine per-instance evidence, steering the model away from spurious token memorisation [Chu et al. 2025] towards principled spatial reasoning. Since recursive partitioning inevitably places objects at crop boundaries, we further introduce a boundary-aware count policy trained via GRPO [Shao et al. 2024] as a post-training objective, which explicitly resolves ownership of straddling instances through nested local, boundary, and global rewards, yielding a substantially more accurate and consistent counter. With the understanding branch thus stabilised, we use it as a frozen counting model to improve the generation branch. Since the two tasks are naturally complementary: understanding maps images to text while generation maps text to images, the understanding branch can directly evaluate how well a generated image matches its text prompt, providing supervision without any external critic. Concretely, we adopt a cycle-consistent GRPO strategy [Shao et al. 2024] where the generation branch produces a group of candidate images for each count-conditioned prompt; the frozen understanding branch scores each candidate based on how well it matches the requested count; an external aesthetic scorer provides a complementary image quality reward; and the combined rewards update only the generation branch. This closed loop progressively improves count-faithful generation without any external model, or human annotation.

In summary, our contributions are as follows:

- We present ABACUS, the first unified VLM that jointly addresses all count understanding tasks like Object counting, Crowd counting, Reference Expression counting and count-accurate image generation in a zero-shot manner.
- We introduce density-aware adaptive zooming paired with an objectness map obtained from MLLM attention layers to spatially ground count predictions, and a novel boundary-aware count policy via GRPO to eliminate the over/undercounting artifact introduced at crop boundaries.
- We propose a cycle-consistent GRPO strategy that uses the frozen understanding branch as an ideal counting model to score generated images via count-deviation and aesthetic rewards, updating

only the generation branch to close the understanding–generation synergy gap.
- ABACUS sets a new state of the art across seven benchmarks spanning object counting, crowd counting, referring-expression counting, count-faithful generation, and count reasoning, surpassing both task-specific specialists and larger generalist models with a single 3B-parameter checkpoint.

2 Related work

2.1 Count Understanding

Existing counting methods are broadly divided into class-specific and class-agnostic approaches. Class-specific methods predict counts for fixed categories such as persons [Guo et al. 2024; Ranasinghe et al. 2024; Shi et al. 2019], vehicles [Hsieh et al. 2017; Mundhenk et al. 2016], and cells [Xie et al. 2018] via detection [Liang et al. 2022; Liu et al. 2023a] or density-map regression [Li et al. 2023a; Wang et al. 2020]. Class-agnostic methods [Chattopadhyay et al. 2017; Lu et al. 2019; Xu et al. 2023a] accept visual exemplars [Liu et al. 2022; Pelhan et al. 2024] or text prompts [Amini-Naieni et al. 2024; Kang et al. 2024; Qian et al. 2025] as targets, though most require dense point-level supervision. Crowd counting [Li et al. 2018; Wang et al. 2020; Zhang et al. 2016a] focuses on dense person scenes via density-map regression or point-level detection [Liu et al. 2023a; Yin et al. 2018], with recent work addressing domain shift [Du et al. 2023; Peng and Chan 2024], uncertainty [Li et al. 2023a], and weakly supervised generalisation [Zhang et al. 2026]. Referring Expression Counting (REC) [Dai et al. 2024b] further generalises class-agnostic counting to free-form expressions requiring joint reasoning over attributes, spatial relations, and category; to date, REC has been addressed exclusively by specialist detection-based models with box-level supervision [Wang et al. 2025a]. ABACUS unifies all three regimes — object counting, crowd counting, and referring-expression counting — within a single zero-shot checkpoint, estimating counts autoregressively via next-token prediction without density maps, counting heads, or point-level annotations, and reports the first unified-VLM result on REC-8K.

2.2 Count Generation

Text-to-image diffusion models [Betker et al. 2023; Black-Forest-Labs 2024; Podell et al. 2024] persistently struggle with accurate object counting [Paiss et al. 2023]. Previous solutions typically rely on architecturally disjoint counting signals, such as contrastive losses [Paiss et al. 2023], auxiliary heads [Kang et al. 2025], or attention manipulation [Chefer et al. 2023]. While spatial planners [Binyamin et al. 2025; Feng et al. 2023; Li et al. 2023b] offer tighter integration, they often fail to generate all specified objects or produce rigid layouts at higher counts [Dahary et al. 2024; Tewel et al. 2024]. Furthermore, recent iterative approaches employing external Vision-Language Model (VLM) critics are inherently bottlenecked by critic reliability [Mondal et al. 2025; Wallace et al. 2024]. To address these limitations, ABACUS unifies the generator, counter, and verifier into a single architecture. By enabling the understanding branch to directly reward the generation branch during training, ABACUS entirely eliminates the need for external critics, planners, or annotations.

2.3 Multimodal Large Models

Early vision-language models (VLMs) such as CLIP [Radford et al. 2021] and BLIP [Li et al. 2022] align vision and text through contrastive pretraining and have been widely adopted for downstream tasks [Jiang et al. 2023b; Kang et al. 2024]. More recently, multimodal large language models (MLLMs) [Bai et al. 2025; Li et al. 2024; Liu et al. 2024a] extend these capabilities with stronger reasoning and generation, achieving notable results on visual question answering [Goyal et al. 2017] and image captioning [Agrawal et al. 2019]. Multimodal Large Language Models (MLLMs) often generate confident yet misaligned outputs [Huang et al. 2024; Sun et al. 2024]. While Reinforcement Learning from Human Feedback (RLHF) effectively mitigates this [Sun et al. 2024; Yu et al. 2024], its reliance on costly manual annotation limits scalability. Consequently, Reinforcement Learning from AI Feedback (RLAIF) has emerged as a practical alternative [Lee et al. 2023; Li et al. 2023c]. Within the counting domain, [Mondal et al. 2025] leveraged AI feedback for count-conditional image generation, employing a critic VLM to iteratively refine outputs without computationally heavy RL training. Conversely, for count understanding, prior works have adapted VLMs for text-driven object counting [Jiang et al. 2023b; Qharabagh et al. 2024]. Methods like CLIP-Count [Jiang et al. 2023b] and VLCounter [Kang et al. 2024] append a specialized counting head to a fine-tuned CLIP backbone. Although WS-COC [Zhang et al. 2026] bypasses explicit heads by autoregressively predicting counts via an MLLM, it remains restricted to weakly-supervised, class-agnostic tasks. In contrast, ABACUS unifies count understanding encompassing object, crowd, and referring-expression counting within a single zero-shot framework. Simultaneously, it enables count-accurate image generation, an unprecedented capability among VLM-based counting methods. We achieve this by fine-tuning our architecture via a novel RLAIF-driven policy optimization.

3 Method

We present ABACUS, a framework that endows a single unified vision-language model (VLM) with both count-accurate image *generation* and precise object *understanding* (counting). We first discuss the preliminaries for RL based model training (Sec 3.1), then we talk about how ABACUS counts object in an image (Sec 3.2), then we discuss how we can generate high cardinality images using the same model (Sec 3.3) and finally, we discuss about the training strategies (Sec 3.4) respectively.

3.1 Preliminaries: Group Relative Policy Optimisation

ABACUS uses Group Relative Policy Optimisation (GRPO) [Shao et al. 2024] as a unified post-training mechanism for both the understanding and generation branches. Unlike PPO, GRPO eliminates the value network by computing advantages relative to a group of rollouts from the policy itself, reducing memory and compute overhead. Given a policy π_θ, an input context $\mathbf{x}$, and a task-specific scalar reward function $R(\cdot)$, GRPO samples K rollouts $\{\mathbf{y}_i\}_{i=1}^{K} \sim \pi_\theta(\cdot \mid \mathbf{x})$

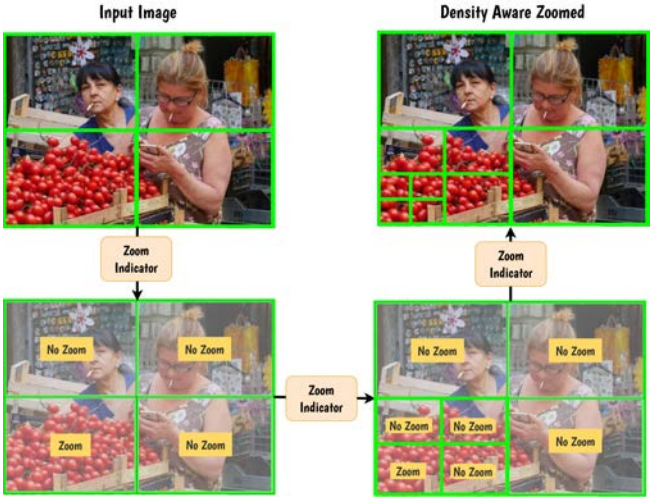

Fig. 3. Density-aware adaptive zooming. The zoom indicator $\phi(\cdot)$ classifies each image region as sparse or dense. Dense regions are recursively partitioned into 2×2 sub-regions until resolution γ is reached; sparse regions are processed in a single pass. Local predictions are aggregated to produce the global count.

and computes group-relative advantages:

$$A_i = \frac{R(\mathbf{y}_i) - \mu_R}{\sigma_R + \epsilon}, \qquad \mu_R = \frac{1}{K}\sum_{j=1}^{K} R(\mathbf{y}_j), \quad \sigma_R = \mathrm{std}\big(\{R(\mathbf{y}_j)\}_{j=1}^{K}\big). \tag{1}$$

The policy is updated by maximising the surrogate objective with a KL penalty against a frozen reference policy π_{ref}:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\sum_{i=1}^{K} A_i \log \pi_\theta(\mathbf{y}_i \mid \mathbf{x})\right] - \beta\, \mathrm{D}_{\text{KL}}\big[\pi_\theta \,\|\, \pi_{\text{ref}}\big], \tag{2}$$

where $\beta > 0$ controls regularisation strength. In section 3, we instantiate this framework twice with different reward functions: a boundary-aware counting reward for the understanding branch (section 3.2) and a count-deviation self-reward for the generation branch (section 3.3).

3.2 Count Understanding

Predicting absolute object counts is challenging for MLLMs, particularly in dense scenes without per-instance supervision. We address this by adaptively partitioning input images into sparser sub-regions, allowing the MLLM to produce more reliable local count estimates that are then aggregated into a global count.

Density-aware Adaptive Zooming Since MLLMs are more reliable on sparse regions, we recursively partition dense images into sub-regions with lower local counts. Partitioning is governed by a zoom indicator $\phi(\cdot)$, implemented as a frozen GroundingDINO [Liu et al. 2024b] backbone with a lightweight learnable MLP head trained on a curated sparse/dense binary dataset. Given an input image I, the indicator computes a density score s_d; if the image is classified as dense, it is recursively split into 2×2 sub-regions until a minimum resolution γ is reached:

$$\{I_i, \ldots, I_n\} = \phi(I; \gamma; s_d), \quad C_i = \mathbb{M}_\theta(I_i; \mathcal{T}), \tag{3}$$

Fig. 4. Infusing objectness in the MLLM. The objectness map is extracted from the attention layers of the language model. Per-head isolation and learned affine alignment produce a spatial distribution over visual token positions, from which object peaks are identified.

where $\mathbb{M}_\theta(\cdot)$ denotes the MLLM head with learnable parameters θ. Following [Tang et al. 2025], we append learnable meta-tokens to the MLLM input. Each sub-image I_i is queried independently to estimate the count of the target specified in text prompt $\mathcal{T}$, and the local predictions C_i are aggregated to produce the final global count.

Infusing Objectness in MLLM Although MLLMs exhibit strong reasoning, they struggle with precise object localisation [Chen et al. 2025a]. Rather than relying on explicit bounding-box prompting, which frequently produces hallucinated coordinates for small objects [Kazemzadeh et al. 2014], we extract spatial object evidence directly from the MLLM's internal representations. Recent work shows that visual tokens largely preserve spatial correspondence to their originating image regions across transformer layers [Neo et al. 2025; Vaswani et al. 2017]; we exploit this by decomposing multi-head self-attention (MHSA) to derive a per-patch objectness signal. Concretely, for each layer l and head i, we isolate the head's contribution via structured masking over the pre-projection concatenated outputs, pass it through the frozen output projection W_O, and apply a shared learned affine alignment $\mathcal{A}(\cdot)$ [Belrose et al. 2023] to obtain a geometrically consistent residual $\mathbf{r}^{l,i}$. The objectness score at each visual token position v is its ℓ_2 magnitude:

$$o^{l,i}(v) = \left\|\mathbf{r}_v^{l,i}\right\|_2, \quad v \in \mathcal{V}. \tag{4}$$

A spatial attention distribution $q^l(v)$ is further derived by averaging cross-attention weights from the final generative token to all visual positions across heads. The objectness map is supervised against Gaussian-smoothed ground-truth point annotations via a objectness regularisation loss:

$$\mathcal{L}_{\text{obj}} = \frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}}\left(-\sum_{v\in\mathcal{V}} g(v)\log\Big(\tilde{q}^l(v) + \epsilon\Big)\right), \tag{5}$$

where $g(v)$ denotes the Gaussian-smoothed ground-truth target, obtained by centring a 2D Gaussian kernel ($\sigma = 20$ px) at each annotated point location and rendering the resulting heatmap at the MLLM attention-map resolution, and $\tilde{q}^l(v)$ is the binarised predicted objectness map. Smoothing converts sparse point supervision into a spatially extended target that matches the continuous space of the attention distribution; aligning the objectness map with these targets therefore encourages the MLLM to internalise instance-aware spatial structure, steering count predictions away from spurious token memorisation and towards principled per-instance spatial reasoning.

Boundary-Aware Count Policy Adaptive zooming inevitably places objects at crop boundaries, causing systematic over- or undercounting [Qharabagh et al. 2024; Zhang et al. 2026]. To resolve

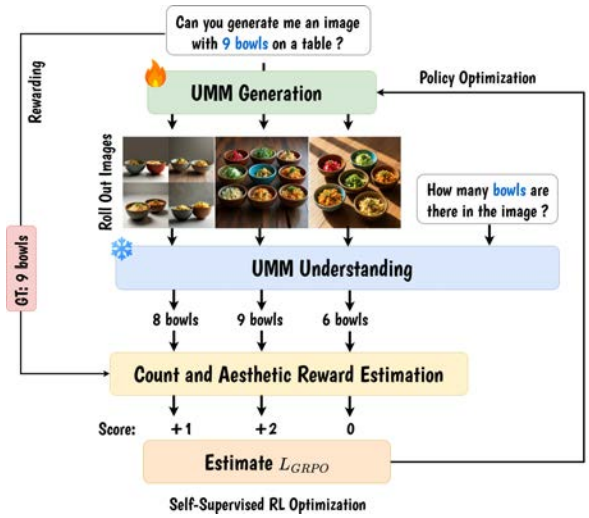

Fig. 5. Count-aware image enhancement via cycle-consistent GRPO. The generation branch samples N candidate images from a count-conditioned prompt. The understanding branch counts objects in each candidate and scores aesthetic quality. The count deviation and aesthetic score form the GRPO reward, which updates the generation branch.

this, we introduce a novel boundary-aware count policy that trains the MLLM to explicitly reason about boundary ownership. Given a dense image partitioned into a 2×2 non-overlapping grid of quadrant crops $\{Q_q\}_{q\in Q}$, where $Q = \{\text{TL}, \text{TR}, \text{BL}, \text{BR}\}$, the policy π_θ produces a structured output $\mathbf{y} = \pi_\theta(\{Q_q\}, \mathcal{P})$ classifying each object as *interior* (interior count n_q^{int}, centroid within Q_q), *edge* (edge-count c_q^e, centroid on this side of cut edge $e \in \mathcal{E}_q$), or *boundary* (boundary-count d_q^e, centroid in the adjacent quadrant). Objects falling exactly on a crop line are assigned to one quadrant randomly, making double-counting structurally impossible. The per-quadrant subtotal s_q and predicted global count $\hat{T}$ are:

$$s_q = n_q^{\text{int}} + \sum_{e \in \mathcal{E}_q} c_q^e, \qquad \hat{T} = \sum_{q \in Q} s_q. \tag{6}$$

The policy is post-trained via GRPO [Shao et al. 2024] using three nested reward components at different spatial granularities: per-quadrant local accuracy (Δ_r^q), cross-quadrant boundary consistency (Δ_r^b), and global count coherence (Δ_r^g), each computed as:

$$\Delta_r = \exp\left(-\frac{|\text{pred} - \text{GT}|}{\text{GT} + \epsilon}\right), \tag{7}$$

which softly penalises over- and undercounting while remaining differentiable, $\epsilon = 1e^{-6}$ preventing division by zero. For each training image, K rollouts $\{\mathbf{y}_i\}_{i=1}^K$ are sampled from π_θ; group-relative advantages are computed as $A_i = (R(\mathbf{y}_i) - \mu_R)/(\sigma_R + \epsilon)$, where μ_R and σ_R are the mean and standard deviation of rewards within the group. The policy is then updated by maximising $\mathcal{L}_{\text{GRPO}}(\theta)$ (using Eq 2). where π_{ref} is a frozen reference policy (the fine-tuned understanding branch checkpoint) and $\beta > 0$ controls the KL penalty strength against reward hacking. The nested rewards jointly drive the policy to resolve boundary ambiguity at every spatial granularity, eliminating the systematic counting errors introduced by adaptive zooming.

3.3 Count Generation

Following UniLIP [Tang et al. 2025], we adopt the SANA diffusion module [Xie et al. 2025], where the autoregressive LM head conditions the diffusion head via cross-attention. Directly using understanding encoder embeddings for generation scrambles spatial layout (fig. 2), revealing a synergy gap from independent branch training, which we address next.

Generation via Understanding Since unified models like UniLIP [Tang et al. 2025] have disentangled autoregressive and diffusion heads with separate training objectives, a synergy gap emerges between the two branches. Yet the tasks are naturally complementary: the understanding branch can directly assess how well a generated image aligns with its conditioning prompt, providing internal feedback without external supervision. Rather than decoupling the two tasks in a "first understand, then generate" pipeline, our approach embeds this feedback directly into generation: the understanding branch identifies failures in the generator's output (*e.g.*, wrong counts or incorrect spatial arrangements) and guides the generation branch to correct them, transforming inter-branch antagonism into a catalyst for progressive generative improvement.

Count Aware Image Enhancement To close the understanding–generation synergy gap, we design a cycle-consistent self-supervised RL framework optimised via GRPO [Shao et al. 2024] (in Eq 2). As illustrated in fig. 5, given a count-conditioned prompt $\mathbf{t}$ specifying target cardinality c^*, the generation branch samples K candidate images $\{\hat{\mathbf{x}}_i\}_{i=1}^K \sim \pi_\theta(\cdot \mid \mathbf{t})$. The frozen understanding branch $\overline{\mathcal{M}}_\theta$ then counts the target instances in each candidate:

$$\hat{c}_i = \text{parse}\Big(\overline{\mathcal{M}}_\theta(\Pi(\mathcal{V}(\hat{\mathbf{x}}_i)),\, \mathbf{z}_t)\Big), \tag{8}$$

yielding a count-deviation reward consistent with eq. (7):

$$R_{\text{cnt}}(\hat{\mathbf{x}}_i) = \exp\left(-\frac{|\hat{c}_i - c^*|}{c^* + \epsilon}\right). \tag{9}$$

To prevent reward hacking through cardinality correctness at the expense of visual quality — the failure mode of attention-manipulation baselines — we augment with an off-the-shelf aesthetic scorer [Schuhmann et al. 2022] $S_{\text{aes}}(\cdot)$, giving the composite reward:

$$R(\hat{\mathbf{x}}_i) = \lambda_c\, R_{\text{cnt}}(\hat{\mathbf{x}}_i) + \lambda_a\, S_{\text{aes}}(\hat{\mathbf{x}}_i), \tag{10}$$

where $\lambda_c = 1.0$ and $\lambda_a = 0.1$ are fixed scalar weights. Group-relative advantages and the KL-regularised surrogate follow eqs. (1) and (2); gradients flow only into the generation-branch LoRA, with π_{ref} fixed to the SFT-initialised generator and the understanding branch frozen throughout. This asymmetry is essential: it prevents the count reward from corrupting the counter that produces it, while exposing the understanding branch to the evolving generator distribution at training time. As the generation branch improves, it produces increasingly realistic multi-instance scenes that progressively sharpen the understanding branch's reward signal, yielding emergent count-faithful synthesis that the generation branch could not achieve when trained in isolation.

3.4 ABACUS Training Strategy

We adopt a three-stage training strategy to build a unified model for count understanding and generation.

Table 1. Comprehensive evaluation of Object Counting (MAE ↓ / RMSE ↓) across FSC-147, CARPK, and ShanghaiTech (SHT), and Count Reasoning (EM ↑) on CountQA. † indicates methods using visual exemplars (few-shot). Methods are partitioned by supervision signal: Point-level (P), Image-level (I), and Zero-shot/MLLM (Z).

Method	Sup.	Object Counting						Crowd Counting				Reasoning
		FSC-147 Val		FSC-147 Test		CARPK Test		SHT-A Test		SHT-B Test		CountQA
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	EM (%) ↑
Specialist Counting Model												
CountGD++ [Amini-Naieni and Zisserman 2025]	P	12.14	47.51	8.39	27.03	–	–	116.0	234.0	28.0	50.0	–
T2ICount [Qian et al. 2025]	P	13.78	58.78	11.76	97.86	8.61	13.47	–	–	–	–	–
CountSE† [Liu et al. 2025]	P	–	–	7.84	82.99	–	–	129.7	258.3	–	–	–
CAD-GD [Wang et al. 2025b]	P	–	–	10.35	86.88	–	–	–	–	–	–	–
VLM-based Counting Model												
GPT-5.5 [OpenAI 2026]	Z	25.87	79.34	25.17	162.0	24.33	36.50	215.67	412.89	58.92	102.34	25.03
Show-o [Xie et al. 2024]	Z	37.87	105.55	46.26	129.53	42.15	63.22	312.45	587.33	89.34	156.78	7.85
Janus Pro 7B [Chen et al. 2025b]	Z	43.56	110.23	35.70	99.96	34.52	51.78	278.91	523.44	76.23	134.56	6.98
UniLIP-3B [Tang et al. 2025]	Z	30.19	103.07	26.44	103.98	26.87	33.48	243.15	424.33	63.79	97.04	9.23
WS-COC-7B [Zhang et al. 2026]	I	14.77	54.24	13.91	97.28	10.39	15.83	128.9	232.9	34.2	57.0	8.44
ABACUS-3B	Z	**5.71**	**26.46**	**5.03**	**27.03**	**8.41**	**10.84**	**78.59**	**139.88**	**14.75**	**25.08**	**15.3**

Stage 1: Understanding Branch Training. We finetune the MLLM using density-aware zoomed images with the combined objective:

$$\mathcal{L}_{\text{und}} = \mathcal{L}_{\text{SFT}} + \mathcal{L}_{\text{obj}}, \tag{11}$$

where $\mathcal{L}_{\text{SFT}}$ is the next-token prediction loss and $\mathcal{L}_{\text{obj}}$ is the objectness regularisation from eq. (5). We train the MLLM, learnable query embeddings, and affine MLP on 2M densely annotated images via LoRA, enabling the model to output structured count predictions. We then apply post-training on 50K curated samples using $\mathcal{L}_{\text{GRPO}}$ to further reduce over/undercounting. The connector and generation branch remain frozen throughout.

Stage 2: Connector Training. The primary focus of this training stage involves optimizing a connector module designed to interface the MLLM with the DiT. By keeping the parameters of both base models frozen and restricting training solely to generative tasks, the connector learns to accurately align the MLLM's output representations with the DiT's original conditional feature space. The connector is trained via SFT ($\mathcal{L}_{SFT}$) with 1M high-quality instruction tuning data respectively. This connector acts as an architectural bridge to transfer the MLLM's inherent world knowledge about count understanding, strong reasoning, and in-context learning capabilities to image generation thus making it count-aware. We direcly feed a set of learnable queries into a frozen MLLM to extract multimodal conditions for multimodal generation which is then passed through this connector into the DiT branch.

Stage 3: Generation Branch Training. We train both the connector and DiT on large-scale count-conditioned generation data, with the MLLM frozen following [Tang et al. 2025]. After SFT, the model can follow generation prompts but produces incorrect cardinalities. We therefore apply post-training via $\mathcal{L}_{\text{GRPO}}$ with combined count-deviation and aesthetic rewards (eq. (10)), updating only the generation branch using count-conditioned prompts as the sole training signal to reduce the generation-understanding synergy.

4 Experiments

4.1 Training Data Collection

Our 2M dense-annotated understanding dataset is curated from large-scale open-source sources including Objects365 [Shao et al. 2019], V3Det [Wang et al. 2023], and SKU-110K [Goldman et al. 2019], retaining only images with a minimum instance count of five objects after numeracy checks, aesthetic filtering, and deduplication. For generation and SFT, we collect 1M images from Pexels, sourced from web alt-text and captions, and filtered via CLIP-based similarity scoring, resolution and aspect-ratio constraints, and text-length checks. Concept-aware sampling is applied to mitigate long-tail distributions, and structured supervision from OCR, charts, and grounding annotations is included to strengthen spatial understanding. These images are annotated with T-Rex2 [Jiang et al. 2023a] for category-labelled boxes, human-verified via T-Rex Label; we prompt Qwen2.5-VL-7B [Bai et al. 2025] with image and verified categories to compose captions. Each prompt's count is the human-verified detector count — grounded, not VLM-estimated — ensuring training-data count accuracy. To ensure fair comparison, all training and test splits of FSC-147 [Ranjan et al. 2021], CARPK [Hsieh et al. 2017], ShanghaiTech [Zhang et al. 2016b], REC [Dai et al. 2024a], and CountQA [Tamarapalli et al. 2025] are strictly excluded from training.

Table 2. Referring expression counting on REC-8K [Dai et al. 2024c] test set (n=3,153 pairs). † uses exemplar images. GroundingREC and finetuned GroundingDino are specialist detection-based methods trained with box-level supervision on REC-8K; ABACUS is a unified VLM evaluated text-only with no training on evaluation benchmark data.

Method	Backbone	FT	MAE ↓	RMSE ↓
Specialist Counting Model				
ZSC [Xu et al. 2023b]	Swin-T	✓	13.00	29.07
TFOC [Shi et al. 2024]	ViT-B	–	17.27	32.68
CounTX [Amini-Naieni et al. 2023b]	ViT-B/16	✓	11.84	25.62
GDino [Liu et al. 2024c]	Swin-T	✓	8.88	21.95
GrREC [Dai et al. 2024c]	Swin-T	✓	**6.50**	19.79
Baselines				
UniLIP-3B [Tang et al. 2025]	–	–	13.75	25.91
ABACUS-3B	UniLIP-3B	–	7.67	**15.84**

4.2 Evaluation Metrics

Following common practice in object and crowd counting [Ranjan et al. 2021; Zhang et al. 2016a], we report Mean Absolute Error (MAE) and, where standard, Root Mean Squared Error (RMSE). For count generation we follow Make-It-Count [Binyamin et al. 2025] and report YOLOv9 [Wang et al. 2024] exact-match accuracy on CoCoCount [Binyamin et al. 2025], the Numeracy score on T2I-CompBench [Binyamin et al. 2025], the Counting subtask of GenEval [Ghosh et al. 2023] averaged over multiple seeds. For referring-expression counting we follow the protocol of GroundingREC [Dai et al. 2024c] and report MAE on REC-8K.

4.3 Implementation Details

Model Architecture and LoRA Configuraion: ABACUS builds on UniLIP-3B [Tang et al. 2025] (InternViT encoder, Qwen2 LLM, SANA-1.5B DiT, DC-AE pixel decoder) connected via a projector Π and N=256 learnable queries. We apply LoRA (r=32, α=64) to all Qwen2 attention (W_Q, W_K, W_V, W_O) and FFN ($W_{\text{up}}, W_{\text{down}}$) projections, adding ~48M trainable parameters (1.6% overhead). The multimodal projector Π and $W_{\texttt{lm\_head}}$ are fully trainable. For text-to-image generation, a separate LoRA (r=16, α=32) is applied to the SANA-1.5B DiT. The visual encoder $\mathcal{V}$, pixel decoder, and query bank remain strictly frozen. The connector consists of 6 layers and maintains the same structure and dimensions as the LLM architecture in InternVL3[Chen et al. 2024].

Post-Training (GRPO): *Boundary Policy* is trained on 8K dense images for 2K steps (LR 5×10^{-6}). We sample K=4 rollouts per image with a strong KL penalty (β=0.04) to stabilize the structured spatial rewards. *Cycle-Consistent Generation* is trained for 5K steps (LR 1×10^{-6}). The frozen understanding branch scores 8 DiT candidates per prompt using reward $r = \exp(-|\hat{c} - c^*|/c^*)$. We use a relaxed β=0.01 to allow synthesis freedom. At inference, this Best-of-8 strategy yields 80% exact numeracy.

Density Indicator & Compute: We use a frozen GroundingDINO-T with a 2-layer MLP classifier trained on 15K sparse/dense images. At inference, regions scoring ≥ 0.5 trigger recursive 2×2 partitioning (max depth γ=3). The entire pipeline is trained in bfloat16 mixed precision, taking ~44 hours on 8× NVIDIA A100 80GB GPUs.

4.4 Comparison with State-of-the-art methods

Object and Crowd Counting: table 1 compares ABACUS against specialist and VLM-based counters. On FSC-147, ABACUS achieves 5.71 val MAE and 5.03 test MAE, surpassing the strongest specialist (CountGD++, 12.14/8.39) by over 40% without point-level annotations. On CARPK, ABACUS attains 8.41 MAE, outperforming the only reporting specialist T2ICount (8.61). On crowd counting, ABACUS achieves 78.59/14.75 MAE on ShanghaiTech-A/B, roughly halving the error of both the best specialist (CountGD++: 116.0/28.0) and the best VLM-based method (WS-COC-7B: 128.9/34.2). Among VLM-based methods, gains over the base UniLIP-3B are 5× on FSC-147 val and 3× on CARPK, confirming that the counting adapter, objectness map, and boundary-aware policy together close the gap to task-specific specialists. Notably, ABACUS is the only method that simultaneously achieves state-of-the-art on both object and crowd counting from a single model. On the count reasoning benchmark CountQA [Tamarapalli et al. 2025], ABACUS achieves 15.3% EM, the highest among open unified VLMs of comparable scale, surpassing UniLIP-3B (9.23%), WS-COC-7B (8.44%), Janus Pro 7B (6.98%), and Show-o (7.85%), respectively.

Table 3. Count Generation evaluation across CoCoCount, T2I-CompBench, and GenEval. YOLOv9 (↑) for CoCoCount and GenEval; Human count accuracy (↑) from annotator study (see supple.); Aesthetic Quality (↑) on GenEval (per-benchmark aesthetics in supplementary).

	CoCoCount		T2I-Comp	GenEval		
Method	YOLOv9	Hum.	Hum.	YOLOv9	Hum.	Aes.
Specialist Count Generation						
CountGen [Binyamin et al. 2025]	50	54	48	46	44	45
BoundedAttn [Dahary et al. 2024]	29	30	35	21	18	10
Count. Guid. [Kang et al. 2025]	21	22	22	16	11	7
VLM-based Count Generation						
BAGEL [Deng et al. 2025]	36	41	32	44	39	43
Janus Pro-7B [Chen et al. 2025b]	27	32	25	30	33	58
UniLIP-3B [Tang et al. 2025]	34	39	30	36	40	61
ABACUS	**71**	**77**	**65**	**94**	**95**	**89**

Referring Expression Counting: On REC-8K [Dai et al. 2024c], ABACUS is queried with free-form referring expressions (*e.g.* "*red apples on the left*") without any architectural modification, achieving MAE 7.67 and RMSE 15.84 over 3,153 evaluation pairs. ABACUS surpasses the fine-tuned GDINO [Liu et al. 2023b] specialist and matches GrREC [Dai et al. 2024c], a detection-trained model with box-level supervision, while achieving substantially lower RMSE (15.84 vs. 19.79). To our knowledge, this is the strongest text-only unified-VLM result on this benchmark.

Count Image Generation: table 3 evaluates count-faithful generation on CoCoCount, T2I-CompBench, and GenEval. ABACUS achieves 71% YOLOv9 exact-match on CoCoCount, surpassing the previous best (CountGen, 50%) by 21 points, and 94 on GenEval counting versus 46 for CountGen. Among unified VLMs, ABACUS nearly doubles the accuracy of the closest competitor BAGEL (36%) with a smaller 3B backbone. Beyond count accuracy, ABACUS achieves the highest aesthetic quality on GenEval (89 vs. 61 for UniLIP-3B), while specialist methods (BoundedAttn, Counting Guidance) score only 7–10 due to attention manipulation and degrading image coherence. Full human evaluation results are in the supplementary, where ABACUS is preferred 39%, 41%, and 50% of the time on CoCoCount, T2I-CompBench, and GenEval respectively, far exceeding the 20% random baseline. An interesting observation of our model is that, during generation branch training, ABACUS relies on the frozen understanding branch, which counts via visible object centers (from objectness regularisation); hence the generation branch thus favours visible centers even under occlusion which we can observe visually in Fig 10(row 5) where our model can generate 28 "colorful balls" with occlusion including the balls underneath which makes it a more realistic generation.

4.5 Ablation Study

We ablate the core components of ABACUS on FSC-147 val (MAE/RMSE) for understanding and CoCoCount (YOLOv9 exact-match) for generation. All variants share the same LoRA adapter, training data, and hyperparameters unless stated otherwise.

Objectness map: We ablate the MHSA-derived objectness map by comparing the full pipeline against: (a) a naive mean-pooled attention map across all heads and layers, and (b) removing objectness regularisation entirely ($\mathcal{L}_{\text{obj}}$=0). As shown in table 4, removing $\mathcal{L}_{\text{obj}}$ causes the largest degradation (+3.92 MAE). Partitioning FSC-147 val into overlap-heavy (≥20% of GT points within 8 px) and overlap-light subsets, the MAE gap between subsets is 2.31 with the full model but widens to 6.18 without $\mathcal{L}_{\text{obj}}$, confirming that the objectness map primarily helps disambiguate spatially proximate instances.

Table 4. Ablation of the objectness map on FSC-147 val.

Variant	MAE ↓	RMSE ↓
Full (per-head + $\mathcal{A}$ + $\mathcal{L}_{\text{obj}}$)	**5.71**	**26.46**
Mean-pooled attention	7.94	34.21
No $\mathcal{L}_{\text{obj}}$	9.63	40.15

Density-aware adaptive zooming: We ablate the zooming module $\phi(\cdot)$ (eq. (3)): (a) no zooming (single-pass), (b) fixed 2×2 grid (unconditional split), and adaptive partitioning (comparing the the objectness map(c) against our G-DINO density indicator(d)) As shown in table 5, single-pass inference suffers on dense scenes. Fixed-grid partitioning improves over single-pass but introduces double-counting at tile boundaries on sparse images, the failure mode that motivates the boundary-aware count policy, ablated in the supplementary. Adaptive zooming avoids both failure modes, yielding the best MAE within 1.2 − 1.3× of single-pass inference. Furthermore, relying on the objectness map instead of a dedicated G-DINO indicator is unreliable for density partitioning, yielding a higher MAE of 8.64 compared to 5.71 for our approach.

Table 5. Ablation of density-aware zooming and indicator choice on FSC-147 val.

Variant	MAE ↓	RMSE ↓	Rel. Time
No zooming (single pass)	10.87	45.32	1.0×
Fixed 2×2 grid	7.93	34.58	∼1.8×
Adaptive (Objectness Map)	8.64	37.21	∼1.3×
Adaptive (G-DINO indicator)	**5.71**	**26.46**	∼1.2×

Count-aware image enhancement: The generation branch is optimised via cycle-consistent GRPO (section 3.3): generate → self-count → compare to prompt → update generator. We compare (table 6): (a) no GRPO (LoRA SFT only), (b) open-loop GRPO (frozen external counter as reward source), and (c) the full cycle-consistent GRPO. SFT alone achieves only 45% exact-match. Open-loop GRPO improves by +17 points, but the full cycle outperforms it by a further 9 points, confirming that co-adaptation of both branches, where the understanding head sharpens the generated images, producing progressively more informative rewards is itself a meaningful source of gain.

Table 6. Ablation of count-aware image enhancement on CoCoCount.

Variant	CoCoCount (exact-match ↑)
No GRPO (LoRA SFT only)	45
Open-loop GRPO (ext. counter)	62
Full cycle-consistent GRPO	**71**

Choice of counting head:: We use the same ABACUS model to count objects via two readouts: (i) *objectness peak*, where the binarised objectness map $\tilde{q}^l(v)$ is thresholded and connected components are counted directly, and (ii) *autoregressive*, where the language head generates the count as text tokens. ABACUS uses the autoregressive readout by default. We compare both from the same model (table 7) with no additional parameters or training. Objectness peak counting is competitive on dense images (≥50 GT objects), where spatial peaks are well-separated, but degrades on sparse images (< 10 GT objects) where the 14×14 patch resolution merges nearby instances and isolated peaks are sensitive to the binarisation threshold. The autoregressive readout outperforms peak counting across both regimes, confirming that $\mathcal{L}_{\text{focus}}$ successfully internalises spatial structure into the language model's hidden states during training, making the explicit spatial readout redundant at inference.

Table 7. Counting using autoregressive head vs. objectness peak on FSC-147 val. Sparse: <10 GT; Dense: ≥50 GT.

	Overall		Sparse	Dense
Readout	MAE ↓	RMSE ↓	MAE ↓	MAE ↓
Objectness peak counting	8.52	36.71	6.31	12.14
Autoregressive (ABACUS)	**5.71**	**26.46**	**3.42**	**9.87**

Training strategies: ABACUS follows a staged initialization plus cycle-consistent GRPO coupling, not a single joint objective. The process fine-tunes UniLIP-3B weights: (i) fine-tune the understanding branch; (ii) freeze it, fine-tune the connector and generation branch on numeracy prompt/image pairs; (iii) GRPO post-training updating only the generation branch using the frozen understanding branch as a reward. As shown in fig. 6, this staged coupling yields complementary gains over either branch trained alone.

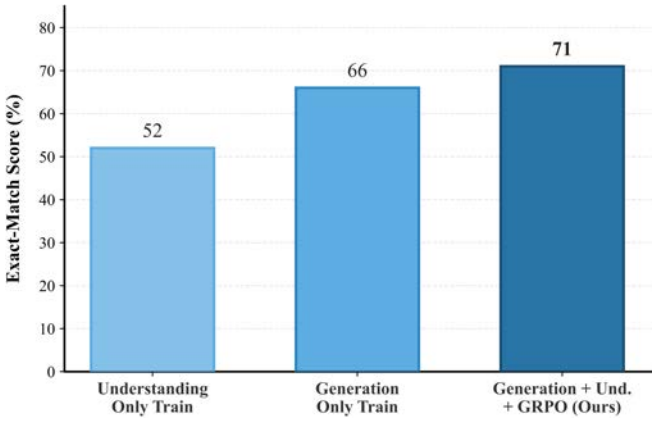

Fig. 6. Ablation of Training Strategy on CoCoCount.

Backbone generalisability: To verify that ABACUS is not specific to UniLIP-3B, we apply the full adapter pipeline to BAGEL-7B [Deng et al. 2025] and Nexus-Gen-7B [Zhang et al. 2025]. As shown in table 8, ABACUS yields consistent gains across all backbones, with performance scaling with model capacity: BAGEL-7B + ABACUS achieves the lowest FSC-147 val MAE (4.93) and highest CoCoCount exact-match (76). This confirms that the adapter pipeline is architecture-agnostic and that scaling the backbone translates directly into stronger counting and generation. All main results are reported on UniLIP-3B to demonstrate state-of-the-art performance at the 3B scale.

High cardinality generation limits: Exact matches remain challenging at extreme densities. As shown in table 9, quality degrades as the requested instance count increases, a physical packing limitation inherent to diffusion models. Nonetheless, ABACUS offers

Table 8. Backbone generalisability of the ABACUS adapter. Larger backbones yield stronger results.

	FSC-147 Val MAE ↓		CoCoCount ↑	
Backbone	Base	+ ABACUS	Base	+ ABACUS
UniLIP-3B	30.19	5.71	34	71
Nexus-Gen-7B	34.81	5.28	32	73
BAGEL-7B	32.45	**4.93**	36	**76**

lower error than task-specific specialists across regimes and operates as a unified open-vocabulary model without task-specific training, unlike density regressors.

Table 9. Count Accuracy and Aesthetics vs. Requested Cardinality.

Count	Accuracy	Aesthetics
10	72.4%	0.91
50	71.1%	0.72
100	68.3%	0.65

Inference cost on dense scenes: Adaptive zooming keeps average inference time within 1.2× of single-pass by partitioning only when the density indicator triggers. For extremely dense scenes requiring maximum recursion depth γ, worst-case latency grows with the number of sub-regions. In practice, such images are rare in standard benchmarks (<3% of FSC-147 val), and the accuracy gains substantially outweigh the cost; nevertheless, early-exit strategies or learned depth prediction could further optimise the speed-accuracy trade-off. We provide a detailed breakdown of inference latency and GPU memory overhead across different density regimes in fig. 7. The overhead is minimal for sparse scenes and scales gracefully up to extremely dense scenes.

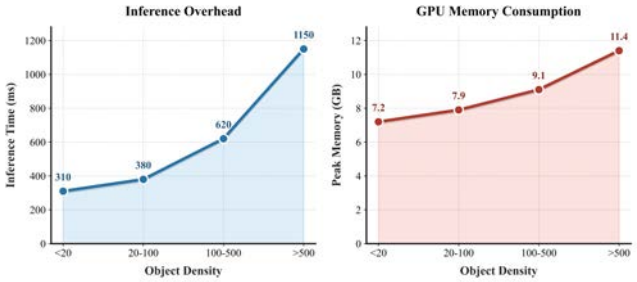

Fig. 7. Inference cost in dense scenes.

5 Limitations and future work

Low-resolution and degraded inputs: ABACUS's counting pipeline relies on spatial tokens from InternViT's 14×14 patch encoding, which requires sufficient input resolution to distinguish individual instances. On low-resolution (<224 px) or heavily compressed images, common in surveillance and legacy datasets, where the visual token grid becomes too coarse for the objectness map to resolve individual objects, a limitation shared by all patch-based VLMs and detection-based counters alike [Amini-Naieni and Zisserman 2025] (see fig. 8. Incorporating resolution-adaptive visual encoders or super-resolution preprocessing could extend ABACUS to these degraded settings.

Single-model generality vs. domain adaptation: A single 3B-parameter model achieves state-of-the-art results across seven benchmarks spanning diverse visual domains. Specialised domains such as medical cell counting [Xie et al. 2018] (see fig. 8), satellite vehicle detection, or industrial defect inspection present distribution shifts that the current count-balanced training mixture does not cover. Lightweight domain adaptation of the LoRA adapter, without retraining the frozen backbone, could unlock these verticals while preserving the general-purpose counting ability, and we leave this exploration to future work.

Fig. 8. Some failure cases of ABACUS.

6 Conclusion

We presented ABACUS, a unified vision-language model for count-aware image understanding and count-faithful generation within a single architecture. Our approach rests on three contributions: an objectness map from MHSA head decomposition that spatially grounds count predictions using point supervision; a novel boundary-aware count policy trained via GRPO with nested rewards that eliminates over/undercounting at crop boundaries; and a cycle-consistent self-reward strategy where the understanding branch counts objects in the generator's own output, closing the feedback loop that external-critic methods leave open. With a single 3B-parameter model, ABACUS sets a new state of the art across object counting, crowd counting, referring-expression counting, and count-faithful generation, surpassing both task-specific specialists and larger generalist models, while establishing the first unified-VLM result on CountQA. These results support a broader justification: count understanding and generation are not competing objectives but mutually reinforcing tasks whose joint optimisation yields emergent spatial awareness that neither specialist alone can achieve, suggesting that the cycle-consistent self-reward paradigm can extend beyond counting to other spatial reasoning tasks where the same model both produces and verifies its outputs.

Acknowledgments

The authors gratefully acknowledge NVIDIA Corporation for support through the NVIDIA Academic Grant Program, which provided computational resources for this research. The authors also acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR) [McIntosh-Smith et al. 2024], operated by the University of Bristol and funded by the UK Government's Department for Science, Innovation and Technology (DSIT) through UK Research and Innovation and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023].

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *ICCV*. 8948–8957.

Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. 2012. Measuring the objectness of image windows. *IEEE transactions on pattern analysis and machine intelligence* 34,

11 (2012), 2189–2202.
N. Amini-Naieni, K. Amini-Naieni, T. Han, and A. Zisserman. 2023a. Open-world Text-specified Object Counting. In *British Machine Vision Conference*.
Niki Amini-Naieni, Kiana Amini-Naieni, Tengda Han, and Andrew Zisserman. 2023b. Open-world text-specified object counting. *arXiv preprint arXiv:2306.01851* (2023).
N. Amini-Naieni, T. Han, and A. Zisserman. 2024. CountGD: Multi-Modal Open-World Counting. In *NeurIPS (NeurIPS)*.
Niki Amini-Naieni and Andrew Zisserman. 2025. CountGD++: Generalized Prompting for Open-World Counting. *arXiv preprint arXiv:2512.23351* (2025).
Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923* (2025).
Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv preprint arXiv:2303.08112* (2023).
James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science* 2, 3 (2023), 8. DALL-E 3 technical report.
Lital Binyamin, Yoad Tewel, Hilit Segev, Eran Hirsch, Royi Rassin, and Gal Chechik. 2025. Make it count: Text-to-image generation with an accurate number of objects. In *CVPR*. 13242–13251.
Black-Forest-Labs. 2024. FLUX. https://github.com/black-forest-labs/flux.
Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R Selvaraju, Dhruv Batra, and Devi Parikh. 2017. Counting everyday objects in everyday scenes. In *CVPR*. 1135–1144.
Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. *ACM Transactions on Graphics (TOG)* 42, 4, 1–10.
Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. 2025a. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In *CVPR*. 513–524.
Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025b. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811* (2025).
Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*. 24185–24198.
Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161* (2025).
Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. 2024. Be Yourself: Bounded Attention for Multi-Subject Text-to-Image Generation. In *ECCV*. 432–448.
Siyang Dai, Jun Liu, and Ngai-Man Cheung. 2024a. Referring Expression Counting. In *CVPR*. 16985–16995.
Siyang Dai, Jun Liu, and Ngai-Man Cheung. 2024b. Referring Expression Counting. In *CVPR (CVPR)*. 16985–16995.
Siyang Dai, Jun Liu, and Ngai-Man Cheung. 2024c. Referring expression counting. In *CVPR*. 16985–16995.
Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. 2025. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683* (2025).
Zhipeng Du, Jiankang Deng, and Miaojing Shi. 2023. Domain-general crowd counting in unseen scenarios. In *AAAI*, Vol. 37. 561–570.
Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. LayoutGPT: Compositional Visual Planning and Generation with Large Language Models. *NeurIPS* 36 (2023), 18225–18250.
Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2023. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS* 36 (2023), 52132–52152.
Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. 2019. Precise Detection in Densely Packed Scenes. In *Proc. Conf. Comput. Vision Pattern Recognition (CVPR)*.
Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*. 6325–6334. doi:10.1109/CVPR.2017.670
Mingyue Guo, Li Yuan, Zhaoyi Yan, Binghui Chen, Yaowei Wang, and Qixiang Ye. 2024. Regressor-Segmenter Mutual Prompt Learning for Crowd Counting. In *CVPR*. 28380–28389.
Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. 2017. Drone-based object counting by spatially regularized regional proposal network. In *ICCV*. 4145–4153.
Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 13418–13427.
Qing Jiang, Feng Li, Tianhe Ren, Shilong Liu, Zhaoyang Zeng, Kent Yu, and Lei Zhang. 2023a. T-rex: Counting by visual prompting. *arXiv preprint arXiv:2311.13596* (2023).
Ruixiang Jiang, Lingbo Liu, and Changwen Chen. 2023b. CLIP-Count: Towards Text-Guided Zero-Shot Object Counting. *arXiv preprint arXiv:2305.07304* (2023).
Seunggu Kang, WonJun Moon, Euiyeon Kim, and Jae-Pil Heo. 2024. VLCounter: Text-Aware Visual Representation for Zero-Shot Object Counting. In *AAAI*, Vol. 38. 2714–2722.
Wonjun Kang, Kevin Galim, Hyung Il Koo, and Nam Ik Cho. 2025. Counting Guidance for High Fidelity Text-to-Image Synthesis. In *WACV*. 899–908.
Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 787–798.
Weicheng Kuo, Bharath Hariharan, and Jitendra Malik. 2015. Deepbox: Learning objectness with convolutional networks. In *ICCV*. 2479–2487.
Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. (2023).
Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326* (2024).
Chen Li, Xiaoling Hu, Shahira Abousamra, and Chao Chen. 2023a. Calibrating uncertainty for semi-supervised crowd counting. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 16685–16695.
Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*. PMLR, 12888–12900.
Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023c. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665* (2023).
Yuheng Li, Haotian Liu, Jianwei Yang, et al. 2023b. GLIGEN: Open-Set Grounded Text-to-Image Generation. In *CVPR*. 22511–22521.
Yuhong Li, Xiaofan Zhang, and Deming Chen. 2018. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*. 1091–1100.
Dingkang Liang, Jiahao Xie, Zhikang Zou, Xiaoqing Ye, Wei Xu, and Xiang Bai. 2023. Crowdclip: Unsupervised crowd counting via vision-language model. In *CVPR*. 2893–2903.
Dingkang Liang, Wei Xu, and Xiang Bai. 2022. An end-to-end transformer model for crowd localization. *ECCV* (2022).
Chengxin Liu, Hao Lu, Zhiguo Cao, and Tongliang Liu. 2023a. Point-query quadtree for crowd counting, localization, and more. In *ICCV*. 1676–1685.
Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. 2022. Countr: Transformer-based generalised visual counting. *arXiv preprint arXiv:2208.13721* (2022).
Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llavanext: Improved reasoning, ocr, and world knowledge.
Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*. Springer, 38–55.
Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*. Springer, 38–55.
Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).
Shuai Liu, Peng Zhang, Shiwei Zhang, and Wei Ke. 2025. CountSE: Soft Exemplar Open-set Object Counting. In *ICCV*. 21536–21546.
Erika Lu, Weidi Xie, and Andrew Zisserman. 2019. Class-agnostic counting. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, 669–684.
Simon McIntosh-Smith, Sadaf R Alam, and Christopher Woods. 2024. Isambard-AI: a leadership class supercomputer optimised specifically for Artificial Intelligence. *arXiv* (2024). https://arxiv.org/abs/2410.11199
Anindya Mondal, Ayan Banerjee, Sauradip Nag, Josep Llados, Xiatian Zhu, and Anjan Dutta. 2025. CountLoop: Training-Free High-Instance Image Generation via Iterative Agent Guidance. *arXiv preprint arXiv:2508.16644* (2025).
T Nathan Mundhenk, Goran Konjevod, Wesam A Sakla, and Kofi Boakye. 2016. A large contextual dataset for classification, detection and counting of cars with deep learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 785–800.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. 2025. Towards interpreting visual information processing in vision-language models. In *ICLR*, Vol. 2025. 57172–57189.

OpenAI. 2026. GPT-5.5 System Card. https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf Accessed: 2026-05-04.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. Teaching clip to count to ten. In *ICCV*. 3170–3180.

Jer Pelhan, Vitjan Zavrtanik, Matej Kristan, et al. 2024. DAVE-A Detect-and-Verify Paradigm for Low-Shot Counting. In *CVPR*. 23293–23302.

Zhuoxuan Peng and S-H Gary Chan. 2024. Single Domain Generalization for Crowd Counting. In *CVPR*. 28025–28034.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *ICLR*.

Muhammad Fetrat Qharabagh, Mohammadreza Ghofrani, and Kimon Fountoulakis. 2024. Lvlm-count: Enhancing the counting ability of large vision-language models. *arXiv preprint arXiv:2412.00686* (2024).

Yifei Qian, Zhongliang Guo, Bowen Deng, Chun Tong Lei, Shuai Zhao, Chun Pong Lau, Xiaopeng Hong, and Michael P Pound. 2025. T2ICount: Enhancing Cross-modal Understanding for Zero-Shot Counting. In *CVPR*.

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

Yasiru Ranasinghe, Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, and Vishal M Patel. 2024. CrowdDiff: Multi-hypothesis Crowd Density Estimation using Diffusion Models. In *CVPR*. 12809–12819.

Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. 2021. Learning to count everything. In *CVPR*. 3394–3403.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems* 35 (2022), 25278–25294.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8430–8439.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).

Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. 2019. Revisiting perspective information for efficient crowd counting. In *CVPR*. 7279–7288.

Zenglin Shi, Ying Sun, and Mengmi Zhang. 2024. Training-free object counting with prompts. In *WACV*. 323–331.

Weibo Shu, Jia Wan, Kay Chen Tan, Sam Kwong, and Antoni B Chan. 2022. Crowd counting in the frequency domain. In *CVPR*. 19618–19627.

Elizabeth S Spelke. 1990. Principles of object perception. *Cognitive science* 14, 1 (1990), 29–56.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the association for computational linguistics: ACL 2024*. 13088–13110.

Jayant Sravan Tamarapalli, Rynaa Grover, Nilay Pande, and Sahiti Yerramilli. 2025. CountQA: How well do MLLMs count in the wild? *arXiv preprint arXiv:2508.06585* (2025).

Hao Tang, Chenwei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. 2025. Unilip: Adapting clip for unified multimodal understanding, generation and editing. *arXiv preprint arXiv:2507.23278* (2025).

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. 2024. Training-Free Consistent Text-to-Image Generation. *ACM TOG* 43, 4 (2024), 1–18.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS* 30 (2017).

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. 2024. Diffusion model alignment using direct preference optimization. In *CVPR*. 8228–8238.

Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. 2020. Distribution matching for crowd counting. *NeurIPS* 33 (2020), 1595–1607.

Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. 2024. Yolov9: Learning what you want to learn using programmable gradient information. In *ECCV*. Springer, 1–21.

Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. 2023. V3det: Vast vocabulary visual detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19844–19854.

Zhicheng Wang, Zhiyu Pan, Zhan Peng, Jian Cheng, Liwen Xiao, Wei Jiang, and Zhiguo Cao. 2025a. Exploring Contextual Attribute Density in Referring Expression Counting. In *CVPR*. 19587–19596.

Zhicheng Wang, Zhiyu Pan, Zhan Peng, Jian Cheng, Liwen Xiao, Wei Jiang, and Zhiguo Cao. 2025b. Exploring contextual attribute density in referring expression counting. In *CVPR*. 19587–19596.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. 2025. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *ICLR*.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528* (2024).

Weidi Xie, J Alison Noble, and Andrew Zisserman. 2018. Microscopy cell counting and detection with fully convolutional regression networks. *Computer methods in biomechanics and biomedical engineering: Imaging & Visualization* 6, 3 (2018), 283–292.

Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. 2023a. Zero-Shot Object Counting. In *CVPR (CVPR)*. 15548–15557.

Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. 2023b. Zero-shot object counting. In *CVPR*. 15548–15557.

Kangxue Yin, Hui Huang, Daniel Cohen-Or, and Hao Zhang. 2018. P2p-net: Bidirectional point displacement net for shape transform. *ACM Transactions on Graphics (ToG)* 37, 4 (2018), 1–13.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13807–13816.

Hong Zhang, Zhongjie Duan, Xingjun Wang, Yuze Zhao, Weiyi Lu, Zhipeng Di, Yixuan Xu, Yingda Chen, and Yu Zhang. 2025. Nexus-Gen: Unified Image Understanding, Generation, and Editing via Prefilled Autoregression in Shared Embedding Space. *arXiv preprint arXiv:2504.21356* (2025).

Xiaowen Zhang, Zijie Yue, Yong Luo, Cairong Zhao, Qijun Chen, and Miaojing Shi. 2026. Bootstrapping MLLM for Weakly-Supervised Class-Agnostic Object Counting. In *ICLR*.

Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. 2016a. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*. 589–597.

Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. 2016b. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*. 589–597.

7 Additional experiments

Fine-grained analysis on REC-8K: To provide further insight into the referring-expression counting capabilities of our model, we present a detailed breakdown of the Mean Absolute Error (MAE) on the REC-8K benchmark. As shown in table 10, we evaluate performance across different expression types, specifically focusing on Attribute and Spatial relationships. ABACUS matches the task-specific specialist GrREC on attribute-based queries and significantly outperforms the base UniLIP-3B backbone across all relationship categories.

Table 10. Detailed MAE breakdown on REC-8K by expression type.

Model	Attribute (MAE ↓)	Spatial (MAE ↓)
GrREC	4.0	**7.5**
UniLIP-3B	8.5	16.0
ABACUS-3B (Ours)	**4.0**	9.0

Choice of reward in boundary policy: Adaptive zooming partitions dense images into quadrant crops, but objects straddling boundaries risk being double-counted or missed. The boundary-aware count policy addresses this via GRPO with three nested rewards: per-quadrant local accuracy (Δ_r^q), cross-quadrant boundary consistency (Δ_r^b), and global count coherence (Δ_r^g). We ablate each independently (table 11). Removing Δ_r^b causes the largest degradation (+1.57 MAE), as the model can no longer arbitrate which quadrant owns a split object. Removing Δ_r^g has a smaller effect (+0.64 MAE): per-quadrant counts are locally accurate but fail to sum coherently. Without GRPO entirely (SFT only), MAE degrades by 2.48, confirming that RL-based reward shaping is necessary. On the dense-activated subset ($\phi(I) = \text{dense}$), the boundary policy improves MAE by 3.74 while leaving sparse-image performance unchanged.

Table 11. Ablation of the boundary-aware count policy on FSC-147 val.

Variant	MAE ↓	RMSE ↓
Full ($\Delta_r^q + \Delta_r^b + \Delta_r^g$)	**5.71**	**26.46**
w/o Δ_r^b (boundary)	7.28	31.84
w/o Δ_r^g (global)	6.35	28.91
w/o Δ_r^q (local)	6.82	30.17
No GRPO (SFT only)	8.19	35.42

8 Human Evaluation

We conduct a human evaluation to assess generated images on *Count Accuracy*, *Aesthetic Quality*, and *Prompt Alignment*, complemented by an overall preference judgment. We recruit 30 annotators via an anonymous Google Form (fig. 11).

Prompt Selection.: We evaluate on a stratified subset of 60 prompts: 25 from CoCoCount, 25 from the T2I-CompBench counting split, and 10 from GenEval, sampled to cover the full count range of each benchmark.

Method Presentation.: We compare seven methods (table 12). Each prompt displays five images labeled A–E: ABACUS is always included, and four competitors are sampled uniformly at random from the remaining six. Labels are shuffled per prompt; annotators are blind to method identity. Under this design, ABACUS appears in all 60 prompt groups per annotator while each competitor appears in ∼40 ($60 \times 4/6$) on average.

Rating Axes.: Each image is rated on a 0–4 Likert scale along three axes:

- **Count Accuracy:** Does the image depict the exact quantity specified in the prompt? 0 = clearly wrong count, 4 = unambiguously correct.
- **Aesthetic Quality:** Visual craftsmanship irrespective of prompt fidelity: composition, clarity, color harmony, and absence of artifacts. 0 = severe distortion, 4 = polished and visually appealing.
- **Prompt Alignment:** Fidelity to the textual prompt *excluding* object count—scene context, spatial arrangement, style, and descriptive details. 0 = major mismatches, 4 = near-perfect alignment.

After rating all five images, the annotator selects a single preferred image considering both prompt alignment and aesthetic quality jointly. **Aggregation and Reporting.:** All Likert scores are normalized to 0–100 (score = $\bar{s}$ / 4 × 100). Per-method scores are averaged over all prompt groups in which that method appears. Preference win rate is the fraction of times a method is selected as the winner among its appearances (random baseline = 20%, since five methods are shown per prompt). Count Accuracy is reported in the main paper; Aesthetic Quality, Prompt Alignment, and Overall Preference are consolidated in table 12 below. For T2I-CompBench, we report human scores only, as YOLOv9 detection is not applicable to the open-vocabulary object classes in this split.

Results.: Table 12 presents the full human evaluation results. ABACUS achieves the highest scores on every axis across all three benchmarks. On aesthetics, ABACUS scores 80, 83, and 89 on CoCoCount, T2I-CompBench, and GenEval respectively, improving over UniLIP-3B (65, 69, 61) despite adding count-specific training, confirming that cycle-consistent GRPO enhances count fidelity without sacrificing visual quality. Specialist methods (BoundedAttn, Counting Guidance) score 7–15 on aesthetics because their attention manipulation degrades image coherence. On prompt alignment, ABACUS leads by a wide margin (79, 79, 90), reflecting its ability to faithfully render scene context and spatial arrangement alongside correct numeracy. Overall preference rates of 39%, 41%, and 50% far exceed the 20% random baseline, with ABACUS preferred 2×–3× more often than the nearest competitor on every benchmark.

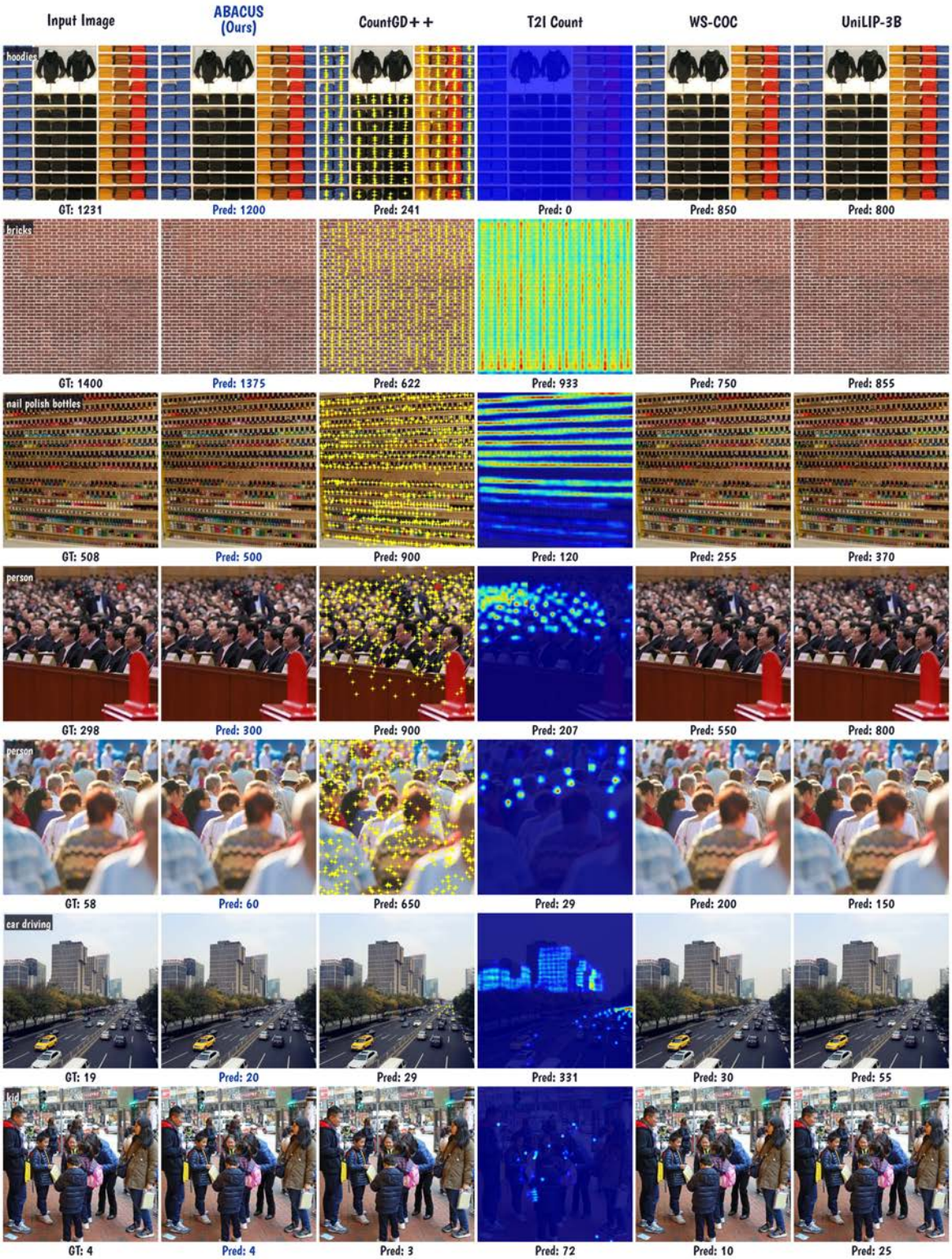

Fig. 9. Qualitative comparison on count understanding. ABACUS (Ours) tracks the ground truth across all density regimes, from sparse (GT: 4) to extremely dense (GT: 1400). CountGD++ overcounts in dense scenes (900 for GT: 298); T2I Count catastrophically fails on out-of-distribution layouts (0 for GT: 1231); WS-COC and UniLIP-3B default to coarse magnitude estimates. Note: In cases of occlusion, counting relies on visible object centers, preventing overcounting from obscured parts.

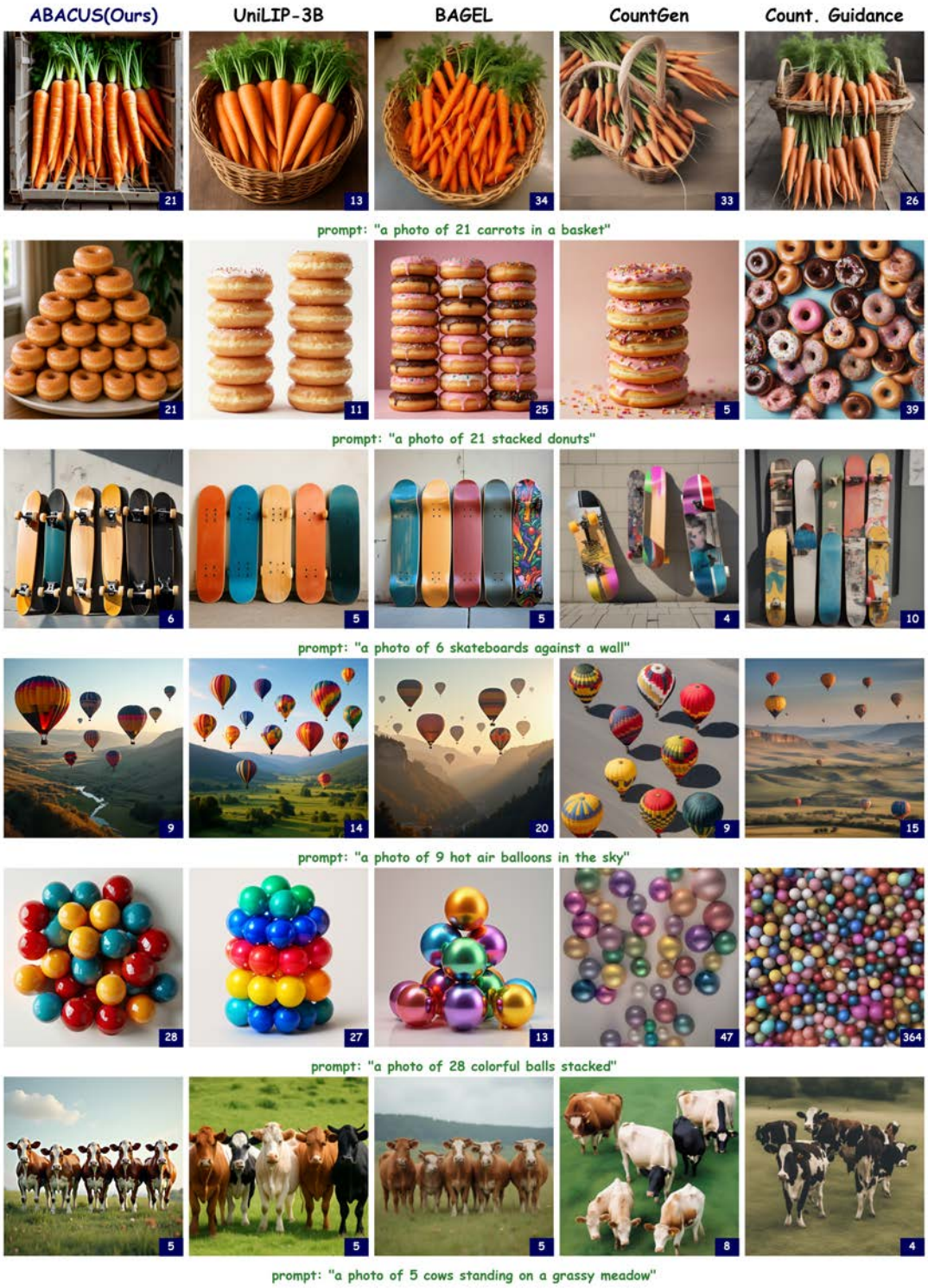

Fig. 10. Qualitative comparison on count generation. ABACUS achieves exact or near-exact counts while preserving natural spatial arrangement. UniLIP-3B systematically undercounts; BAGEL overcounts with unnatural compositions; CountGen produces rigid grid patterns; Counting Guidance exhibits mode collapse (39 donuts for a prompt of 21, 364 balls for a prompt of 28).

Table 12. Full human evaluation results: Aesthetic Quality (Aesth. ↑), Prompt Alignment (Align. ↑), and Overall Preference (Pref. ↑, win rate %). All Likert scores normalized to 0–100. Preference random baseline is 20%.

Method	CoCoCount			T2I-CompBench			GenEval		
	Aesth. ↑	Align. ↑	Pref. (%) ↑	Aesth. ↑	Align. ↑	Pref. (%) ↑	Aesth. ↑	Align. ↑	Pref. (%) ↑
Specialist Count Generation									
CountGen [Binyamin et al. 2025]	45	47	15	50	50	18	45	45	12
BoundedAttn [Dahary et al. 2024]	11	15	2	8	13	2	10	12	1
Counting Guidance [Kang et al. 2025]	15	16	2	10	12	2	7	8	1
VLM-based Count Generation									
BAGEL [Deng et al. 2025]	55	52	14	48	45	12	43	42	10
Janus Pro-7B [Chen et al. 2025b]	60	54	12	52	47	10	58	53	11
UniLIP-3B [Tang et al. 2025]	65	60	16	69	61	15	61	57	15
ABACUS	**80**	**79**	**39**	**83**	**79**	**41**	**89**	**90**	**50**

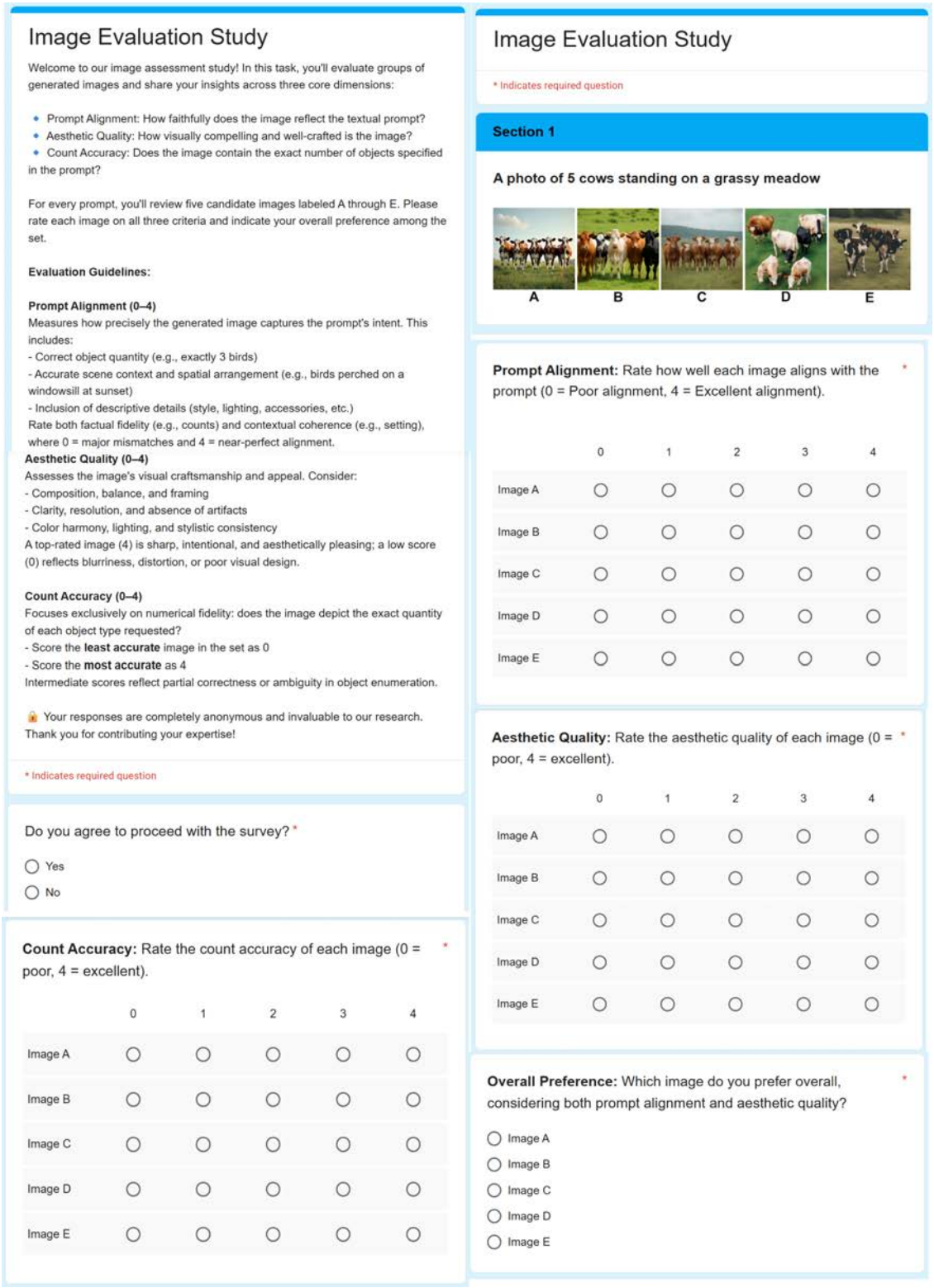

Fig. 11. Human evaluation form. For each prompt, annotators rate five anonymized images on Count Accuracy, Aesthetic Quality, and Prompt Alignment (0–4 Likert), then select an overall preferred image.

Fig. 12. Count understanding gallery. ABACUS predictions (green) across diverse categories and count ranges from FSC-147, CARPK, and ShanghaiTech. The model achieves exact or near-exact counts from sparse scenes (GT: 1, pencil) to dense crowds (GT: 261, go stones) using text-only prompts.

Fig. 13. Count generation gallery. ABACUS generates images with the exact requested count across diverse prompts, maintaining naturalistic spatial arrangement and high aesthetic quality.