Occlusion-Aware Text-To-Image Generation

Bach-Hoang Ngo[1,2,3†], **Si-Tri Ngo**[1,2†], **Hieu Le**[3‡], **Trung-Nghia Le**[1,2‡*]

[1]University of Science, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]UNC Charlotte, North Carolina, United States

https://bachngoh.github.io/AIBL/

Abstract

Text-to-image diffusion models fail to generate correct object counts in dense scenes, where overlapping instances collapse into indistinguishable structures despite appearing visually plausible. We identify this as instance ownership collapse: tokens from overlapping objects interact freely through attention, while heavily occluded instances receive weak supervision due to their small visible areas. We address this through layout-aware attention biases that softly bias token interactions toward region-consistent grouping and suppress cross-instance leakage, paired with an amodal-balanced loss that amplifies gradients for occluded objects based on their occlusion level. To enable systematic evaluation, we introduce OverlapDepth-45K, a benchmark of densely overlapping scenes with amodal supervision. Our approach substantially improves count accuracy and prevents instance merging while preserving image quality.

1 Introduction

Generating correct content is a fundamental requirement for controllable media generation[30, 35, 34, 21, 14, 40, 22, 37, 38]. Yet modern text-to-image (T2I) models remain unreliable when asked to produce scenes with explicit object counts [21, 2, 30]. Prompts specifying dozens of instances frequently result in missing objects, merged entities, or visually plausible but numerically incorrect compositions. As shown in Figure 1, while a standard model handles simple prompts (3 swans), it overcounts single-category scenes (7 instead of 6 apples), drops instances in multi-category compositions (7 instead of 12 objects), and catastrophically merges dense scenes (50+ instead of 28 across persons and cans). This limitation significantly hinders the utility of T2I models in applications demanding high structural fidelity.

Free-form generation without structural priors rarely satisfies cardinality constraints. Conditional generation with layout or mask priors partially alleviates this problem by anchoring objects spatially [25, 43]. However, once scenes become dense and objects heavily overlap, failures persist. Even with correct bounding boxes, instances collapse into indistinguishable blobs: the model may allocate roughly the correct semantic mass within a region, but individual objects are not rendered as separable, countable entities.

We argue that this failure arises from how diffusion transformers represent and learn overlapping objects. In dense layouts, tokens from different instances frequently occupy nearby or overlapping spatial regions. Standard self-attention allows features to freely propagate across these regions, causing representations of adjacent objects to blend together during generation. Meanwhile, the diffusion training objective applies a uniform reconstruction loss over all pixels, which implicitly

†Equal contributions ‡Equal advising *Corresponding author
Emails: nhbach22@apcs.fitus.edu.vn, ntsi22@apcs.fitus.edu.vn, hle40@charlotte.edu, ltnghia@fit.hcmus.edu.vn

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

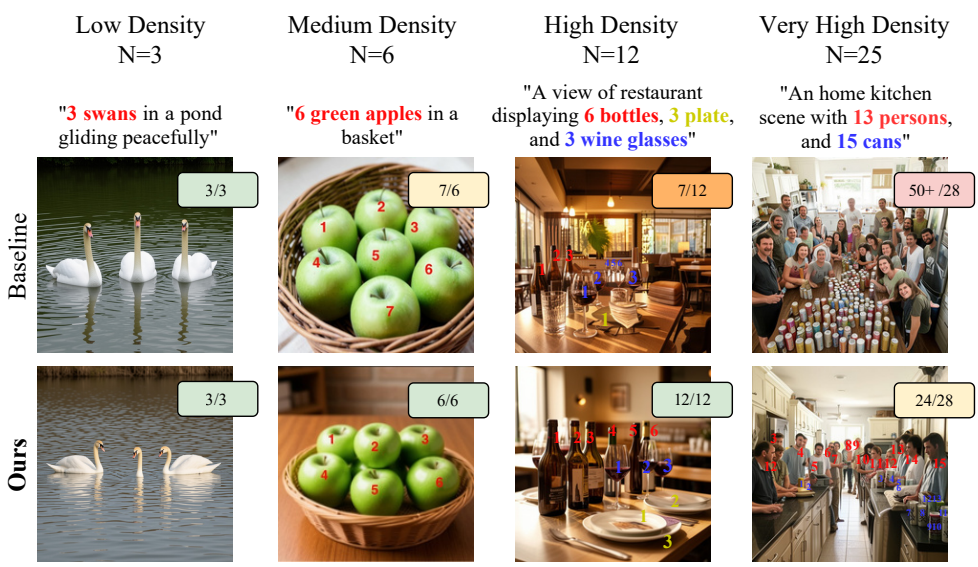


Figure 1: **Count-preserving multi-category compositional generation.** Standard diffusion models (top) succeed on simple prompts (3/3 swans) but fail as scene complexity grows - overcounting single-category scenes (7/6 apples), dropping instances in multi-category compositions (7/12 across bottles, plates, and glasses), and catastrophically merging objects in dense scenes (50+/28 across persons and cans). Our approach (bottom) achieves exact or near-exact counts across all density levels (3/3, 6/6, 12/12, 24/28) while preserving photorealistic quality and natural inter-object occlusion.

favors large and fully visible regions. When objects are heavily occluded, only a small fraction of their pixels contributes to the loss, producing weak gradients for those instances. Over training, this imbalance teaches the model to allocate most of its capacity to visible regions while neglecting partially occluded objects. Consequently, when layouts overlap, the model tends to synthesize shared textures across neighboring boxes rather than maintaining distinct instance boundaries.

To address this, we introduce a framework that encourages instance ownership throughout generation and training. We inject layout-aware attention biases into the transformer to softly favor region-consistent token grouping, while simultaneously applying an amodal-balanced loss that amplifies gradients for occluded instances based on the ratio between their full and visible extents. These two mechanisms operate together: layout attention improves instance separability in dense overlaps, and the balanced objective ensures that occluded objects remain strong optimization targets during training.

Concretely, our *layout-aware attention* modifies the attention logits of the diffusion transformer using region identities derived from the input layout. Tokens belonging to the same bounding box receive a positive cohesion bias that encourages intra-instance interaction, while tokens from different boxes receive a separation bias that suppresses cross-instance feature leakage. This simple modification provides soft spatial guidance to reduce local blending when regions overlap, rather than enforcing hard geometric adherence to box boundaries. Complementing this structural constraint, the *amodal-balanced loss* adjusts the diffusion training objective to account for occlusion. During training, we estimate each instance's occlusion level from its amodal and visible masks and compute a scaling factor proportional to the ratio between its full extent and visible area. This factor is applied to the reconstruction loss of pixels owned by the instance, ensuring that partially visible objects contribute gradients comparable to fully visible ones.

To systematically study dense compositional generation under severe overlap, we further introduce **OverlapDepth-45K**, a dataset of highly populated scenes with scalable amodal supervision derived from standard instance annotations. This dataset provides controlled evaluation of object counting and instance separability across a wide range of occlusion levels.

In summary, our key contributions are:

1. **Ownership-Aware Layout Attention:** We introduce a parameter-free attention bias for DiT architectures that exploits the depth-ordered ownership structure of overlapping bounding boxes. By reinforcing cohesion for ownership-winner tokens in contested zones and strengthening separation between overlapping instances, the mechanism prevents cross-instance feature leakage without rigid hard-masking or additional trainable parameters.
2. **Amodal-Aware Instance-Balanced Loss (AIBL):** We propose a novel training-time loss reweighting scheme that amplifies gradients for occluded instances using amodal-to-visible area ratios, improving counting fidelity in dense scenes without introducing inference-time mask dependencies.
3. **Empirical Validation:** We demonstrate that layout attention combined with AIBL effectively mitigates catastrophic object fusion, reducing RMSE by 35.3% (from 9.26 to 5.99) over the layout-only baseline on our OverlapDepth-45K benchmark while maintaining high precision on tolerance thresholds. Our proposed approach significantly outperforms existing methods on T2I-Compbench [15] and CoCoCount [3] datasets.

2 Related Work

Counting in T2I. While counting is a classic problem in computer vision[31, 11, 32, 41, 39, 10], modern text-to-image diffusion models still struggle with precise counting, especially as object counts increase [15]. Existing solutions include test-time guidance from auxiliary counting networks [17], resampling-based correction methods[24] such as CountGen [3], and training losses that encourage stronger token-object alignment [6]. These methods improve sparse scenes but often scale poorly to dense layouts. A key limitation is that most counting or attention losses aggregate activations globally or inside coarse boxes, without distinguishing visible from occluded object regions. As a result, overlapping objects can be incorrectly penalized as missing, causing unstable gradients in dense scenes. We instead introduce an amodal-aware objective that reweights instance contributions by visibility, improving high-density counting without expensive test-time optimization.

Layout-to-image generation and spatial control. Layout-to-image methods reduce prompt ambiguity by conditioning generation on spatial priors. GLIGEN [25], ControlNet [42], T2I-Adapter [28], LayoutDiffusion [43], and later methods such as InstanceDiffusion [33], MIGC [44], and 3DIS-FLUX [45] improve object placement through grounded tokens, attention constraints, decoupled decoding, or depth-aware control. Training-free layout guidance has also been explored through cross-attention manipulation [7, 29]. However, layout conditioning mainly provides a soft placement prior. Under severe box overlap, standard layout guidance can still collapse multiple instances into one blended object. In our framework, layout attention anchors object locations, while the proposed amodal-aware loss directly addresses the dense instance-fusion problem.

Occlusions and dense overlaps. Severe occlusion remains difficult for compositional generation. SeeThrough3D [1] models occlusion-aware 3D boxes, while OMG [19] uses a two-stage layout and comprehension pipeline to reduce missing entities. Recent benchmarks such as OverLayBench [23] further show that amodal masks alone are insufficient for dense overlaps. The core issue is not only knowing what object shapes exist, but deciding how training gradients should be assigned when instances overlap. Our Amodal-Aware Instance-Balanced Loss (AIBL) addresses this optimization imbalance by weighting instances according to occlusion degree, leading to stronger count preservation in highly dense scenes.

3 Proposed Method

Generating correct object counts in dense scenes fails not because diffusion models lack spatial awareness, but because their training objectives are blind to occlusion. Under standard uniform reconstruction losses, heavily occluded instances contribute negligible gradients, effectively teaching the model to drop them. Layout conditioning alone cannot fix this: even with correct bounding boxes, the optimization imbalance persists.

We address this with two complementary components (see Figure 2): (1) a **Layout Attention mechanism** that injects region-aware spatial biases into the transformer's attention logits, softly encouraging region-consistent interactions at both training and inference without additional trainable parameters; and (2) the **Amodal-Aware Instance-Balanced Loss (AIBL)**, which scales each instance's loss contribution in inverse proportion to its visibility, restoring balanced gradient flow without introducing any inference-time mask dependency.

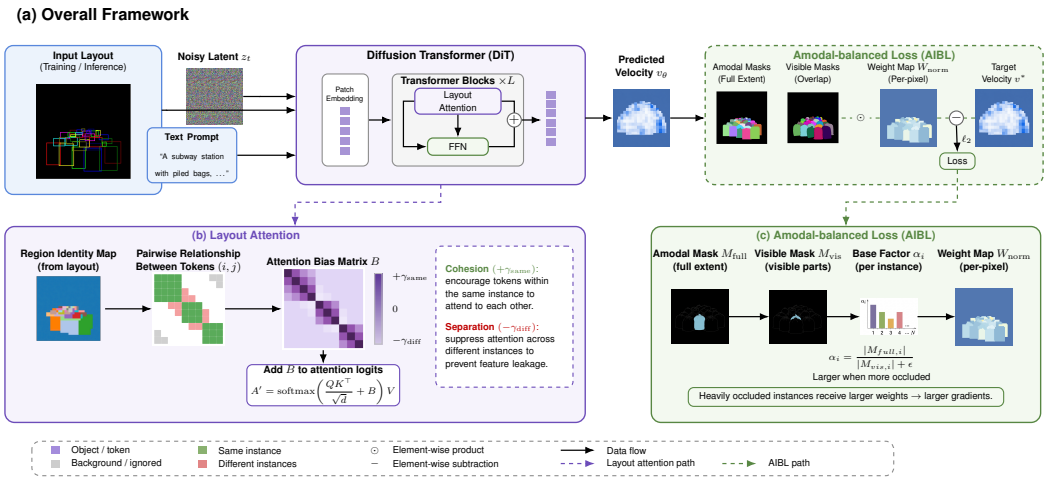


Figure 2: **Architecture Overview.** (a) Layout Attention (LA) is integrated into the transformer blocks of a DiT backbone to steer the generation process. (b) LA utilizes a region identity map to compute an attention bias matrix M, where cohesion and separation biases are applied to tokens based on their instance ownership. (c) The Amodal-Aware Instance-Balanced Loss (AIBL) module (training only) computes per-instance occlusion boost factors α_i from amodal and visible masks. These factors produce a mean-normalized weight map W_{norm} that prioritizes heavily occluded objects during training.

3.1 Layout Attention

Rather than introducing separate trainable cross-attention modules or manipulating latents at inference via gradient guidance, we inject soft region priors directly within the existing attention mechanism of the diffusion transformer. Our layout attention requires no additional learnable parameters beyond the LoRA adapters and operates identically at training and inference, permanently closing any training-inference gap.

Region identity map. Given a text prompt $\mathcal{P}$ with N target instances, we define a layout as $\mathcal{B} = \{b_1, b_2, \ldots, b_N\}$, where each $b_i = (x_i, y_i, w_i, h_i) \in [0, 1]^4$ specifies the normalized spatial coordinates for the i-th instance. We construct a discrete *region identity map* $\mathcal{R} \in \mathbb{Z}^{H_l \times W_l}$ over the latent spatial grid by painting bounding boxes in depth order (i.e., a painter's algorithm): each token is assigned the region ID of the *frontmost* instance whose box covers it, and background tokens receive an ID of 0. This depth-ordered assignment is the key structural prior that our attention bias exploits.

Overlap-ownership attention bias. In dense layouts, bounding boxes inevitably overlap. When two boxes intersect, the painter's algorithm assigns contested tokens exclusively to the depth-front instance—the *ownership winner*. The occluded instance retains only its non-overlapping tokens, which may be a small fraction of its total extent. This asymmetry creates a natural *ownership structure* over the latent grid: tokens in overlapping regions belong unambiguously to a single instance, but they are spatially adjacent to tokens of the occluded object and thus especially prone to cross-instance feature leakage through attention.

We exploit this structure by computing a per-token *coverage count* $c(p) = |\{b_i \in \mathcal{B} \mid p \in b_i\}|$ and defining the overlap zone $\mathcal{Z} = \{p \mid c(p) \geq 2\}$. The attention bias matrix $M \in \mathbb{R}^{S \times S}$ (where $S = S_{\text{txt}} + H_l W_l$) over all pairs of image tokens is then:

$$M(p_i, p_j) = \begin{cases} \beta \cdot \omega_{\text{intra}} + \Delta_{\text{coh}}(p_i, p_j) & \text{if } \mathcal{R}(p_i) = \mathcal{R}(p_j) > 0 \\ \omega_{\text{inter}} + \Delta_{\text{sep}}(p_i, p_j) & \text{if } \mathcal{R}(p_i) \neq \mathcal{R}(p_j), \text{ both} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $\omega_{\text{intra}} > 0$ is the base intra-instance cohesion bias, $\omega_{\text{inter}} < 0$ is the base cross-instance separation bias, and $\beta > 1$ is a global grounding factor for dense scenes. The overlap-adaptive

corrections Δ_{coh} and Δ_{sep} activate only in the presence of contested tokens. Let $\mathcal{Z}_k = \{p \mid \mathcal{R}(p) = k,\ p \in \mathcal{Z}\}$ denote the set of overlap-zone tokens owned by instance k. The corrections are:

$$\Delta_{\text{coh}}(p_i, p_j) = \begin{cases} \frac{1}{2}\beta\omega_{\text{intra}} & p_i, p_j \in \mathcal{Z}_k \text{ for some } k, \\ 0 & \text{otherwise,} \end{cases} \qquad \Delta_{\text{sep}}(p_i, p_j) = \begin{cases} \frac{1}{2}\omega_{\text{inter}} & \mathcal{Z}_{\mathcal{R}(p_i)} \neq \emptyset \text{ or } \mathcal{Z}_{\mathcal{R}(p_j)} \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The cohesion correction Δ_{coh} reinforces the identity of ownership-winner tokens that lie in the contested zone, preventing their features from being "pulled" toward the occluded instance. The separation correction Δ_{sep} operates at the *region* level: once any instance has tokens in a contested area, *all* cross-instance pairs involving that instance receive stronger repulsion. This whole-region propagation is deliberate—it prevents leakage not just at the overlap boundary but along the entire interface between neighboring instances, which is where instance merging initiates. Both corrections scale the existing bias magnitudes by a fixed factor of $\frac{1}{2}$, which we do not tune across experiments.

Gaussian soft prior. To complement the discrete region bias with smooth spatial guidance, we augment M with a distance-weighted Gaussian centered at each bounding box's centroid:

$$M_{\text{soft}}(p_j, b_i) = \exp\left(-\frac{\|\text{pos}(p_j) - \text{center}(b_i)\|^2}{2\sigma^2}\right). \tag{3}$$

This provides a gentle spatial decay that softens the hard region boundaries, encouraging tokens near a box's center to attend more strongly to their assigned instance while gradually releasing peripheral tokens.

The combined bias $M + M_{\text{soft}}$ is injected as an additive term into the pre-softmax attention logits of the DiT's joint attention blocks. Because this modifies only the attention distribution rather than introducing new trainable pathways, layout attention preserves the pretrained generative prior while providing ownership-aware soft region guidance for count-faithful instance separation.

3.2 Amodal-Aware Instance-Balanced Loss (AIBL)

Standard diffusion models minimize a uniform Mean Squared Error (MSE) over all spatial positions equally. This creates a severe optimization imbalance in occluded objects. Consider a training image containing two objects: a large foreground sofa occupying 40% of the image and a heavily occluded cushion with only 5% of its area visible. The sofa generates roughly $8\times$ more gradient signal than the cushion under uniform MSE, simply because it occupies more pixels. Over training, the model learns that reconstructing the occluded object has little impact on the loss. As a result, occluded instances are often merged or omitted at inference.

Occlusion boost factor. AIBL addresses this by leveraging amodal masks *exclusively to reweigh the loss function*, ensuring the model receives amplified error signals for occluded instances without conditioning on masks at inference. For each instance i in a training sample, we compute a multiplicative occlusion boost factor:

$$\alpha_i = \min\left(\alpha_{max}, \left(\frac{A_i^{amodal}}{A_i^{visible} + \epsilon}\right)^{\gamma}\right), \tag{4}$$

where A_i^{amodal} is the instance's theoretical full (unoccluded) area, $A_i^{visible}$ is its actually visible area in the composite, $\gamma \in (0, 1)$ is a controlling parameter, and α_{max} clips extreme values. Intuitively, a fully visible object receives $\alpha_i \approx 1$, while a 90%-occluded object receives $\alpha_i \gg 1$.

Pixel-wise weight map construction. Instance-level boost factors must be translated into a spatial weight map that can be applied to the per-pixel reconstruction loss. We construct $W \in \mathbb{R}^{H \times W}$ by assigning each pixel the boost factor α_i of its owning instance, where ownership is determined by the depth-ordered ownership map $\mathcal{O}$ provided by the dataset (i.e., the frontmost instance at each pixel is its owner):

$$W(x, y) = \begin{cases} \alpha_{\mathcal{O}(x,y)} & \text{if } \mathcal{O}(x, y) > 0 \\ \lambda_{bg} & \text{if } \mathcal{O}(x, y) = 0 \end{cases} \tag{5}$$

where $\mathcal{O}(x, y) \in \{0, 1, \ldots, N\}$ denotes the instance ID of the frontmost object at pixel (x, y), with 0 reserved for background. Background pixels receive a downweighted baseline value $\lambda_{bg} = 0.8$, reflecting that background regions carry less instance-specific information and should not dominate the loss even when they are large.

Mean normalization. Applying per-instance boosts naively would inflate the overall loss magnitude, implicitly increasing the effective learning rate in dense scenes. To prevent this, we mean-normalize W across all spatial positions within each training sample:

$$W_{norm}(x, y) = \frac{W(x, y)}{\frac{1}{HW} \sum_{x', y'} W(x', y')}. \tag{6}$$

This ensures that the sum of all pixel weights equals HW fore every training image, exactly preserving the global gradient scale regardless of scene complexity. A heavily occluded scene and a simple scene therefore produce gradients of the same expected magnitude, and the learning rate requires no adjustment.

3.3 Training Objective

The complete architecture is trained end-to-end using rectified flow matching [27]. Given a clean latent $\mathcal{Z}_0$, a noise sample ϵ, and a random timestep $t \sim \mathcal{U}(0, 1)$, the noisy latent is formed as $\mathcal{Z}_t = (1 - t)\mathcal{Z}_0 + t\epsilon$ with target velocity $v_{target} = \epsilon - \mathcal{Z}_0$. The AIBL-weighted training objective is:

$$\mathcal{L}_{total} = \mathbb{E}_{\mathcal{Z}_0, \epsilon, t, \mathcal{P}, \mathcal{B}} \left[W_{norm} \odot \| v_\theta(\mathcal{Z}_t, \mathcal{P}, \mathcal{B}) - v_{target} \|_2^2 \right], \tag{7}$$

where v_θ denotes the model's velocity prediction conditioned on the layout-biased attention. To match the diffusion latent space, the pixel-space amodal and ownership masks are downsampled to the latent resolution via interpolation prior to computing the weight map W_{norm}. When AIBL is disabled, $W_{norm} = \mathbf{1}$ and the objective reduces to standard uniform MSE. Layout attention bias M (Eq. 1) is applied identically during both training and inference, ensuring no training-inference gap.

4 Experiments

Implementation Details. We fine-tune FLUX.2-Klein-base-4B with rank-16 LoRA for 3,000 steps at learning rate 5×10^{-5}. We use $(\omega_{\text{intra}}, \omega_{\text{inter}}, \beta) = (2, -1, 2)$, $\sigma = \sqrt{0.05}$, $\gamma = 0.5$, $\alpha_{\max} = 5$, and $\lambda_{\text{bg}} = 0.8$, selected on a separate 50-image development set and then fixed. Appendix F gives parameter roles, the attention-strength sweep, and protocol details.

4.1 OverlapDepth-45K Dataset

Training a depth-aware, layout-guided model requires a large-scale dataset featuring reliable amodal bounding boxes and pairwise ordinal depth metadata, particularly in highly occluded scenarios. Existing datasets like MS-COCO [26] or Visual Genome [20] lack the dense, strictly controlled occlusion complexity required to elicit overlap ambiguity. To address this, we introduce the *OverlapDepth-45K* dataset.

Our OverlapDepth-45K dataset comprises 45,000 images, split into 44,000 training images and 1,000 test images. It features densely populated scenes with an average of 13.2 objects per image (typically ranging from 6 to 20 objects) - see Figure 3. To the best of our knowledge, it is the most challenging dataset for high-object-count scene generation, particularly due to the high object density and substantial occlusion. It spans 92 object categories, including people and common household items, across diverse environments such as offices, streets, and supermarkets, etc. The dataset intentionally captures severe visual ambiguity, with an average object occlusion rate of 26.6%. An example is shown in Figure 4.

4.2 Evaluation Protocol and Metrics

We evaluate our model's counting fidelity using a held-out test set of densely populated prompts. The generated images are subjected to an automated strict evaluation protocol leveraging zero-shot

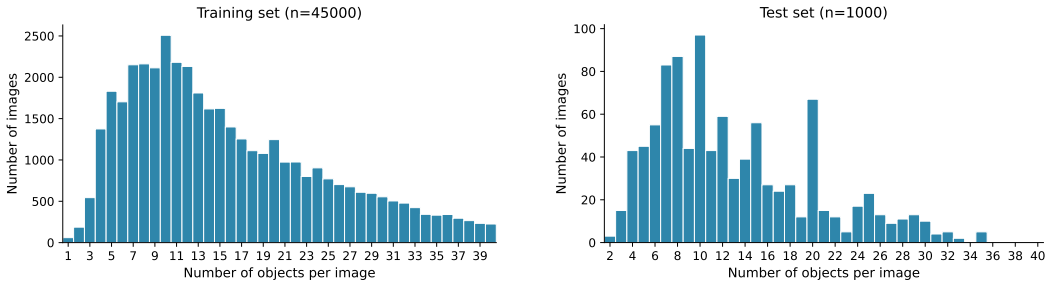


Figure 3: Histogram showing the proportion of images with respect to the number of objects in the training set (left) and test set (right). The distribution highlights the dataset's focus on dense scenarios.



Figure 4: Visual sample from our training dataset. We show the original image, visible mask overlays, amodal mask overlays, ownership mask for depth ordering, and the revised prompt.

detection pipelines (SAM 3 [5]) tuned for high precision. To ensure the reliability of this automated evaluation and mitigate potential biases, we validate the SAM 3 detections against manual human annotations on a subset of the data. Our primary metrics include: **Exact Match Accuracy (%):** The percentage of object categories where the predicted count exactly matches the target count. **Relative Tolerance Accuracy ($\pm 10\%, \pm 20\%$):** The percentage of object categories whose predicted count falls within a strict percentage-based error margin, scaling naturally with large object counts. **Error Magnitude (MAE and RMSE):** Mean Absolute Error measures the average absolute deviation between target and detected counts per category, while Root Mean Square Error disproportionately penalizes massive outlier failures (such as the fusion of many objects).

4.3 Comparison Against State-of-the-Arts

Table 1: **OverlapDepth-45K** (1,000-image test): T2I vs. L2I under a fair layout protocol. Layout rows use identical **LLM-inferred** boxes and labels generated by Gemini 3 Flash (Section 4.3).

		Accuracy (%) ↑			Error ↓	
Method	**Type**	Exact	$\pm 10\%$	$\pm 20\%$	MAE	RMSE
FLUX.2-Klein-4b (text only)	T2I	13.24	18.18	30.59	5.21	9.96
SDXL-Turbo [30] (text only)	T2I	5.30	9.50	18.50	10.05	14.71
GLIGEN [25]	L2I	8.84	15.48	29.15	11.41	18.76
MIGC [44]	L2I	10.46	19.12	30.17	4.83	6.96
3DIS [45]	L2I	12.88	19.48	30.17	4.07	6.34
AIBL (Ours)	**L2I**	**15.32**	**20.12**	**33.11**	**3.89**	**5.99**

Text-to-image vs. layout-to-image baselines on OverlapDepth-45K. Benchmarking mixes two confounds: whether a model receives explicit spatial anchors, and *which* boxes are supplied at test time. To ensure a fair comparison and separate paradigm differences from layout-quality bias, Table 1 contrasts **text-to-image (T2I)** systems that only observe the revised count prompts against **layout-to-image (L2I)** systems that additionally receive the same *LLM-inferred layout* boxes (generated via Gemini 3 Flash) and category labels. This matches spatial conditioning across GLIGEN [25],

MIGC [44], and 3DIS [45], and uses the identical setup for our model. Automated counting follows Section 4.2. As shown in Table 1, our method consistently outperforms all L2I baselines across all metrics, yielding an absolute Exact Match improvement of +2.44% over the strongest baseline (3DIS) and reducing RMSE from 6.34 to 5.99.

Table 2: State-of-the-art comparison on CoCoCount and the T2I-Compbench counting subset. Best results are highlighted in **bold**.

	CoCoCount [3]		**T2I-Compbench** [15]
Method	**YOLOv9 Acc.**	**Human Acc.**	**Human Acc.**
SDXL [30]	28	26	29
Repeated Object [12]	17	18	14
Reason Out Your Layout [8]	21	26	15
DALL-E 3 [2]	25	38	36
Random masks + BoundedAttn [29]	29	30	35
Counting Guidance [16]	21	22	22
CountGen (Make-It-Count) [3]	50	54	48
AIBL (Ours)	**58**	**60**	**65**

Comparison on external benchmarks. To further validate the effectiveness of our approach, we evaluate our method on the counting subsets of T2I-Compbench [15] and CoCoCount [3] against state-of-the-art methods, including recent specialized counting pipelines like CountGen [3]. As shown in Table 2, we report baseline numbers as provided by Binyamin et al. [3]. Our trained model significantly outperforms existing approaches. We randomly sample a subset of 100 images for human evaluation on T2I-Compbench. Under this evaluation, our approach achieves 65% accuracy, establishing a massive absolute improvement of +17% over CountGen (48%) and +29% over DALL-E 3 (36%). This robust performance on complex, open-world prompts demonstrates that the principles of amodal gradient weighting and instance isolation generalize beyond our dense synthetic distributions.



Figure 5: **Qualitative comparison on counting prompts.** Even with explicitly grounded bounding boxes, the layout-attention baseline (Base) frequently merges nearby instances. Our Amodal-Aware Instance-Balanced Loss (AIBL) rebalances training to promote clearer instance separation, producing more distinct objects (Ours). We also include reference generations from SDXL, FLUX.1-dev [21], and Z-Image [4] on the same prompts.

Qualitative Evaluation We visually compare generations from the layout attention baseline against our Layout Attn. + AIBL model in Figure 5. The baseline exhibits pronounced "feature diffusion" in dense regions - fusing canine bodies into a monolithic mass (*11 dogs*) or merging pots into indistinct green blobs (*16 plants*) -rather than generating discrete objects. In contrast, AIBL restores strict instance boundary separation: because heavily occluded regions receive proportionally stronger training gradients, the model preserves structural integrity even in constrained areas, resolving overlapping bottles, plants, and animals as countable entities.

4.4 Ablation Study

We conduct ablation studies on the OverlapDepth-45K test set to isolate the contribution of each component. We evaluate the following configurations: (1) **Untrained (FLUX.2)**: the unconditioned

baseline without layout training, (2) **Basic Layout Attn.**: layout attention without overlap-ownership bias, (3) **Basic Layout Attn. + AIBL (γ=0.5)**: adding occlusion-aware loss reweighting, and (4) **Ours (Full)**: our complete model adding the overlap-ownership bias. All trained models use identical LoRA rank, learning rate (5×10^{-5}), and training duration (3000 steps).

Table 3: Counting accuracy on the OverlapDepth-45K test set. Exact denotes Exact Match Accuracy while $\pm 10\%$ and $\pm 20\%$ denote Relative Tolerance Accuracy. Best results are highlighted in **bold**.

	Accuracy (%) ↑			Error ↓	
Model	Exact	$\pm 10\%$	$\pm 20\%$	MAE	RMSE
Baseline	13.24	18.18	30.59	5.21	9.96
Basic Layout Attn.	14.71	18.24	31.18	4.47	9.26
+ AIBL (γ=0.5)	**15.34**	19.21	32.12	4.22	8.12
+ Overlap-Ownership (Ours)	15.32	**20.12**	**33.11**	**3.89**	**5.99**

Effectiveness of components. As shown in Table 3, integrating basic layout attention substantially improves counting over the untrained baseline. Adding AIBL further amplifies these gains. Finally, applying our Overlap-Ownership bias (γ=0.5) achieves the strongest results on nearly every metric, with RMSE dropping to 5.99 - a massive 35.3% reduction from the basic layout-only baseline (9.26). Because RMSE penalizes large-magnitude errors quadratically, this reduction provides compelling evidence that AIBL mitigates *catastrophic failure modes*: the severe object merging in dense layouts that causes massive undercounting spikes. Crucially, this configuration improves nearly *all* metrics simultaneously, demonstrating that occlusion-aware reweighting not only eliminates catastrophic failures but also provides a broadly beneficial training signal.

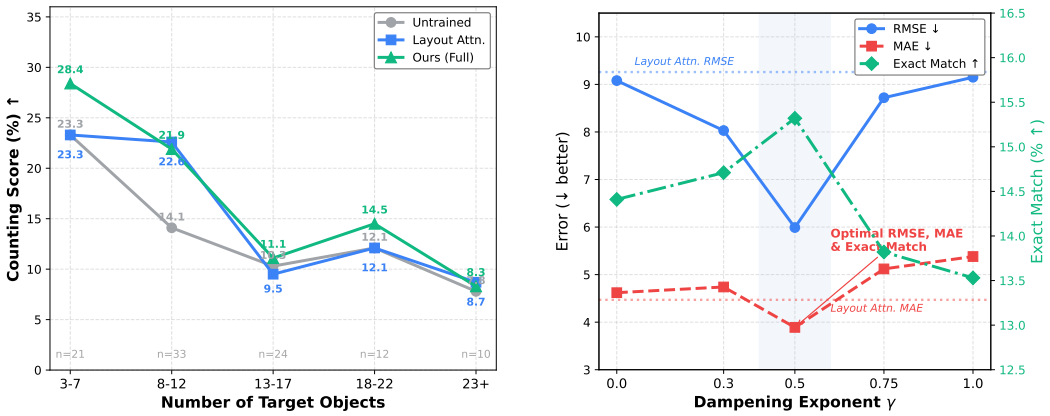
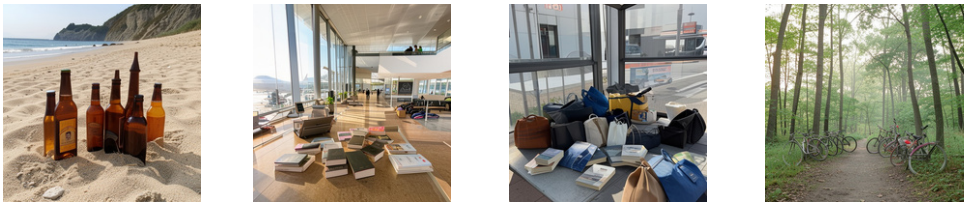


Figure 6: **Analysis of AIBL behavior.** (*Left*) Counting score stratified by target object density. AIBL provides significant improvements in mid-to-high density regimes (13–22 objects). (*Right*) RMSE, MAE, and Exact Match accuracy as a function of the dampening exponent γ. Dotted lines indicate the layout-only baseline. The optimal γ=0.5 achieves the best trade-off across all metrics.

Counting Score by Scene Density. We further evaluate the *counting score* - the percentage of target objects correctly generated and detected - stratified by scene density (Figure 6a). Unlike binary Exact Match, this metric captures partial success (e.g., 18/20 cups = 90%). AIBL provides its strongest improvements in the mid-to-high density regimes (13 - 22 objects), where occlusion is most severe: in the 18 - 22 bin it achieves 14.5% vs. 12.1% for layout-only (19.8% relative gain), confirming that gradient rebalancing is most impactful where occluded instances are most likely to be dropped. Even in the low-density regime (3 - 7 objects), AIBL delivers a 5.1 percentage-point gain, suggesting that amodal-aware weighting helps preserve instance boundaries across densities. In the highest bin (23+ objects) the layout-only baseline marginally outperforms AIBL (8.7% vs. 8.3%), likely because extreme overlap exceeds the regime where gradient boosting alone can resolve all instances - an open challenge for future work.

Sensitivity to dampening exponent γ. The dampening exponent γ in Equation (4) controls the aggressiveness of occlusion-based reweighting: smaller values compress the boost range, while larger values preserve the raw amodal-to-visible ratio but risk gradient instability. We evaluate

$\gamma \in \{0.0, 0.3, 0.5, 0.75, 1.0\}$ with all other hyperparameters fixed. Results reveal a clear inverted-U pattern (Figure 6b). At γ=0, AIBL reduces to foreground-vs-background reweighting, yielding only a modest RMSE improvement (9.08 vs. 9.26 for layout-only). The optimal γ=0.5 achieves the best results across nearly *all* metrics simultaneously, striking the ideal balance between gradient amplification and training stability. Beyond this point, performance degrades: at γ=1.0, a 90%-occluded object receives a 10$\times$ raw boost, and the resulting RMSE (9.15) approaches the layout-only baseline, effectively negating the benefits of occlusion-aware reweighting.



Prompt: "An authentic sandy beach scene with 5 bottles, naturally lit."

Prompt: "A wide-angle view of airport terminal displaying 9 books, clear natural light."

Prompt: "An authentic bus stop shelter scene with 10 books, 2 umbrellas, and 12 bags, clear natural light."

Prompt: "An authentic forest trail scene with 7 bicycles, soft ambient lighting."

Figure 7: **Failure cases.** Our method can struggle with extreme density or thin objects.

4.5 Discussion and Limitations

While AIBL effectively mitigates object fusion, three inherent limitations remain (Figure 7). (1) *Density saturation*: In extremely packed scenes, latent space resolution limits force feature overlap, causing the model to merge nearby objects despite balanced gradients. (2) *Category bias*: Thin or articulated objects (e.g., umbrellas) lose structural detail during latent compression, making them harder to render than simpler shapes (e.g., bags), inflating per-category errors. (3) *Synthetic domain shift*: OverlapDepth-45K is entirely synthetic. While zero-shot evaluations on real datasets (e.g., T2I-CompBench) show strong generalization, synthetic scenes can encode non-canonical lighting and artifactual layouts. Future work should explore explicit shape priors, category-aware loss reweighting, and real-world hybrid datasets to overcome these challenges.

5 Conclusion

In this paper, we introduced Amodal-Aware Instance-Balanced Loss (AIBL), a novel training-time reweighting scheme that dynamically scales the loss contribution of each instance based on the ratio of its amodal extent to its visible area. Coupled with a distance-weighted grounded layout attention mechanism and our dedicated OverlapDepth-45K benchmark, AIBL amplifies supervisory gradients for heavily occluded objects without introducing inference-time mask dependencies or training-inference mismatch. Experimental results show that our proposed method effectively mitigates catastrophic object fusion, while strictly preserving baseline photorealism and image quality. Future work will focus on overcoming this extreme-density overlapping objects, potentially drawing inspiration from spatial priors that explicitly reason about 3D occlusion to better disentangle complex semantic overlaps.

Acknowledgments

This research is funded by Vietnam National University - Ho Chi Minh City (VNU-HCM) under Grant Number B2026-18-17 and the faculty startup fund from the College of Computing and Informatics (CCI) at the University of North Carolina at Charlotte.

References

[1] Vaibhav Agrawal, Rishubh Parihar, Pradhaan Bhat, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu. Seethrough3d: Occlusion aware 3d control in text-to-image generation. *arXiv preprint arXiv:2602.23359*, 2026.

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[3] 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)*, 2025.

[4] Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

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

[6] 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), July 2023.

[7] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[8] Xiaohui Chen, Yongfei Liu, Yingxiang Yang, Jianbo Yuan, Quanzeng You, Li-Ping Liu, and Hongxia Yang. Reason out your layout: Evoking the layout master from large language models for text-to-image synthesis. *arXiv preprint arXiv:2311.17126*, 2023.

[9] diffusers. Stable diffusion xl 1.0 inpainting. `https://huggingface.co/diffusers/stable-diffusion-xl-1.0-inpainting-0.1`, 2023.

[10] Corentin Dumery, Noa Ette, Aoxiang Fan, Ren Li, Jingyi Xu, Hieu Le, and Pascal Fua. Counting stacked objects. In *IEEE/CVF International Conference on Computer Vision*, 2025.

[11] Dimitri Gominski, Maurice Mugabowindekwe, Qiue Xu, Xiaowei Tong, Martin Brandt, Hieu Le, Rasmus Fenshol, Dimitris Samaras, and Loic Landrieu. Counting trees from satellite imagery with noisy supervision. In *European Conference on Computer Vision*, volume 17032, pages 296–313, 2026.

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

[13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.

[14] Shilin Hu, Jingyi Xu, Akshat Dave, Dimitris Samaras, and Hieu Le. Physics-grounded shadow generation from monocular 3d geometry priors and approximate light direction, 2026. Preprint.

[15] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

[16] Wonjun Kang, Kevin Galim, Hyung Il Koo, and Nam Ik Cho. Counting guidance for high fidelity text-to-image synthesis. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 899–908. IEEE, 2025.

[17] Wonjun Kang and Kevin Park. Counting guidance for high fidelity text-to-image synthesis. *arXiv preprint arXiv:2306.17567*.

[18] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021.

[19] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXI*, page 253–270, Berlin, Heidelberg, 2024. Springer-Verlag.

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[21] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[22] Hieu Le, Jingyi Xu, Nicolas Talabot, Jiancheng Yang, and Pascal Fua. Pairwise-constrained implicit functions for 3d human heart modeling. In *Medical Image Computing and Computer Assisted Intervention*, 2025.

[23] Bingnan Li, Chen-Yu Wang, Haiyang Xu, Xiang Zhang, Ethan J. Armand, Divyansh Srivastava, Xiaojun Shan, Zeyuan Chen, Jianwen Xie, and Zhuowen Tu. Overlaybench: A benchmark for layout-to-image generation with dense overlaps. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[24] Shuangqi Li, Hieu Le, Jingyi Xu, and Mathieu Salzmann. Enhancing compositional text-to-image generation with reliable random seeds. In *International Conference on Learning Representations*, 2025.

[25] 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 *CVPR*, 2023.

[26] 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*. Springer, 2014.

[27] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023.

[28] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304, 2024.

[29] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7932–7942, 2024.

[30] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

[31] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *European Conference on Computer Vision*, 2018.

[32] Kostas Triaridis, Panagiotis Kaliosis, E-Ro Nguyen, Jingyi Xu, Dimitris Samaras, and Hieu Le. What is the right embedding space for contrastive learning in referring expression counting? In *European Conference on Computer Vision*, 2026.

[33] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6232–6242, 2024.

[34] Haoyu Wu, Jingyi Xu, Qiaomu Miao, Dimitris Samaras, and Hieu Le. Phase-aligned rope for mixed-resolution diffusion transformer. In *European Conference on Computer Vision*, volume 17005, pages 272–293, 2026.

[35] Haoyu Wu, Jingyi Xu, Dimitris Samaras, and Hieu Le. Importance-based token merging for efficient image and video generation. In *IEEE/CVF International Conference on Computer Vision*, 2025.

[36] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 15903–15935, 2023.

[37] Jingyi Xu and Hieu Le. Generating representative samples for few-shot classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[38] Jingyi Xu, Hieu Le, Mingzhen Huang, ShahRukh Athar, and Dimitris Samaras. Variational feature disentangling for fine-grained few-shot classification. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[39] Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. Zero-shot object counting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[40] Jingyi Xu, Hieu Le, and Dimitris Samaras. Assessing sample quality via the latent space of generative models. In *European Conference on Computer Vision*, 2024.

[41] Jingyi Xu, Hieu Le, and Dimitris Samaras. Learning to count from pseudo-labeled segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025.

[42] 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*, 2023.

[43] 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, June 2023.

[44] 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*, pages 6818–6828, 2024.

[45] Dewei Zhou, Ji Xie, Zongxin Yang, and Yi Yang. 3DIS: Depth-driven decoupled image synthesis for universal multi-instance generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

This supplementary material provides additional details and experimental analyses supporting our proposed method. We first present reproducibility and ethical considerations in Appendix A. We then describe the OverlapDepth-45K dataset (Section A) and implementation details (Section B). Additional qualitative results and ablations are provided in Sections C and D, followed by human validation of automated counting metrics (Section E). We further report additional experimental details (Section F), image-fidelity evaluation (Section G), fine-grained AIBL ablations (Section H), and an expanded human-preference study (Section I).

Appendix A: Preliminaries and Declarations

A.1 Reproducibility Statement

To ensure the reproducibility of our proposed approach, we provide the implementation details in Section B. Our evaluations are conducted on the publicly available T2I-Compbench and CoCoCount datasets and our proposed OverlapDepth-45K dataset. The proposed OverlapDepth-45K dataset and source code of the proposed AIBL will be made publicly available upon publication.

A.2 Ethics and Societal Impact

AIBL inherently carries risks if misused to create deceptive content. To mitigate these risks, we advocate for and utilize digital watermarking techniques (e.g., SynthID) to ensure provenance. We strictly adhere to the Toronto Declaration; this framework is designed solely for creative and educational purposes and must not be used for malicious profiling or discrimination.

A.3 Human Subjects and Asset Disclosures

Our human evaluation ($N = 5$) was granted an Institutional Review Board (IRB) exemption. All adult participants provided informed consent, and no personally identifiable information (PII) was collected.

A OverlapDepth-45K Dataset

In this section, we provide further details regarding our OverlapDepth-45K dataset, specifically focusing on the distribution of object counts, the diverse category groupings, and visual examples of our comprehensive annotations.

Category Groupings Our dataset spans diverse real-world scenarios such as "*Office*", "*Street*", or "*Living Room"*, as summarized in Table 4.

Dataset Generation Pipeline The construction of OverlapDepth-45K follows a multi-stage automated pipeline designed to ensure both scene density and annotation precision (Figure 8). First, we utilize a Large Language Model (LLM) to generate complex counting prompts and corresponding spatial layouts (bounding boxes) for diverse scene groups. Second, these layouts and prompts are passed to a high-capacity text-to-image model (FLUX.2-Klein-9B) to synthesize the base images. Third, we employ SAM 3 to extract precise visible instance masks for each object anchored to the layout. Finally, our amodal expander (described below) reconstructs the occluded regions of each object to produce full amodal masks and depth-ordered ownership maps. This pipeline allows for the scalable creation of dense compositional scenes with reliable ground truth signals for occlusion-aware training.

Amodal Inpainting Details To generate the amodal masks, we employ the Stable Diffusion XL 1.0 Inpainting model [9]. For a given visible instance mask (extracted via SAM 3), we mask out the occluding foreground objects and prompt the inpainter with the target category (e.g., "a complete, unoccluded bottle"). The inpainted completion is then re-segmented by SAM 3 to obtain the full amodal extent of the object. To quantitatively validate the reliability of this automated amodal expander, we conducted a manual visual inspection of 100 randomly sampled amodal masks from our generated dataset. Human evaluators verified that over 94% of the generated amodal completions

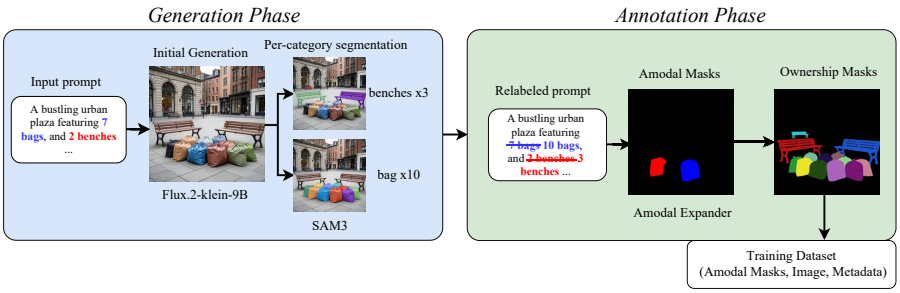


Figure 8: **Dataset Generation Pipeline.** Our automated pipeline for constructing the OverlapDepth-45K dataset involves: (1) LLM-based layout and prompt synthesis, (2) high-fidelity image generation, and (3) automated annotation of visible and amodal masks via SAM 3 and category-conditioned inpainting.

Table 4: List of object categories organized by scene groups.

Scene Group	Object Categories
Office/Indoor	laptop, monitor, keyboard, mouse, tv, book, pen, clock, telephone, remote, handbag, backpack
Street/Outdoor	car, bus, train, motorcycle, bicycle, airplane, boat, traffic light, stop sign, fire hydrant, parking meter
Living/Furniture	couch, chair, dining table, bed, bench, toilet, sink
Kitchenware	refrigerator, oven, microwave, toaster, cup, bowl, plate, wine glass, fork, knife, spoon, bottle
Animals/Pets	dog, cat, bear, zebra, horse, cow, sheep, elephant, giraffe, bird
Foods/Fruits	apple, banana, orange, carrot, broccoli, pizza, cake, donut, hot dog, sandwich, pastry
Sports/Toys/Misc	skis, snowboard, tennis racket, skateboard, surfboard, frisbee, baseball bat, baseball glove, sports ball, teddy bear, kite, umbrella, vase, potted plant, scissors, hair drier, toothbrush, candle, pill, key, coin, wallet, box, blanket, pillow, towel

were physically plausible and geometrically consistent with the visible regions. This confirms that our Amodal-Aware Instance-Balanced Loss (AIBL) is supervised by robust, high-quality occlusion signals.

Visualization. Figure 9 visualizes a few training examples from our dataset. For each sample, we provide the corresponding *amodal masks* (capturing the full extent of objects regardless of occlusion), *ownership masks* (capturing the depth-ordered visible regions), and the *revised prompt* describing the layout structure.

B Implementation Details

B.1 Computing Resources

Experiments are conducted on a single NVIDIA A100 GPU.

B.2 Implementation

We train our models on the OverlapDepth-45K dataset, which contains 44,000 dense training images and 1,000 held-out test images used exclusively for evaluation.

Our proposed layout attention is injected into the DiT joint attention blocks of FLUX.2-Klein-4b. All variants are fine-tuned via LoRA ($r = 16$) at a learning rate of 5×10^{-5}, converging after 3,000 iterations in mixed precision.

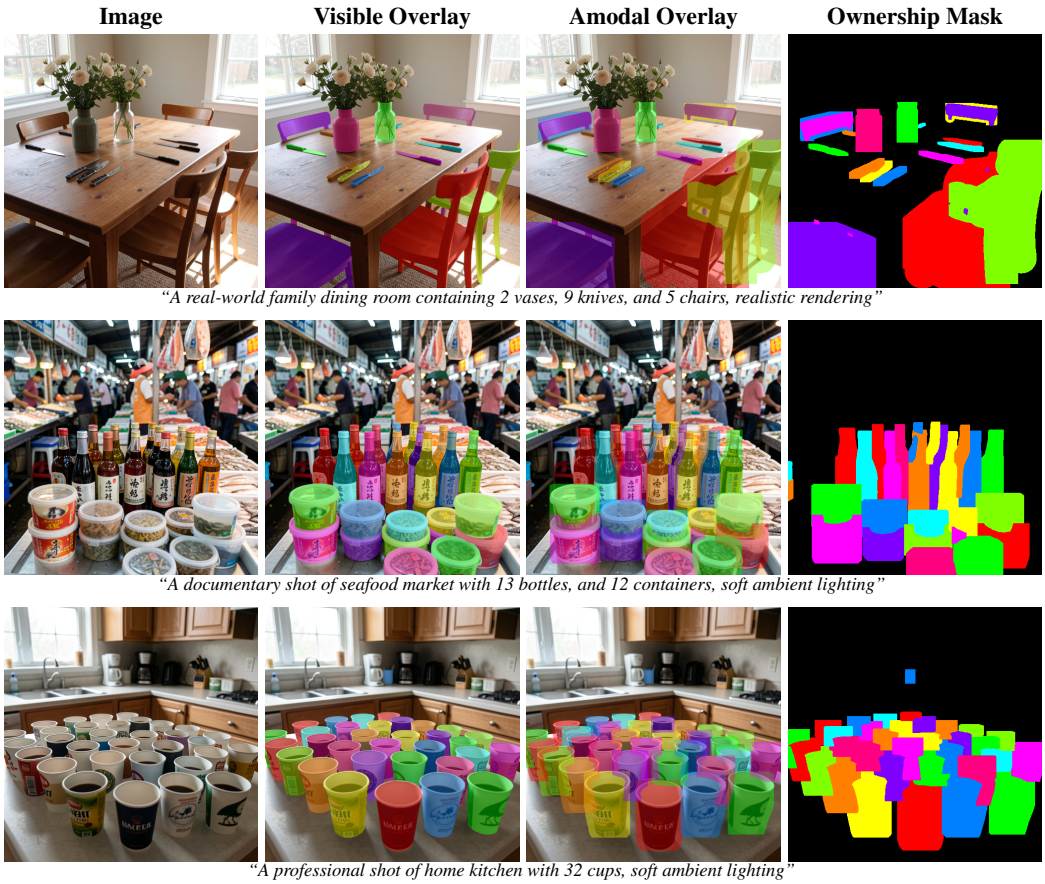


Figure 9: Visual samples from our training dataset. For each sample, we provide: (1) the original image, (2) visible mask overlays showing unoccluded regions, (3) amodal mask overlays capturing full object extents, (4) the ownership mask for depth ordering, and the revised prompt.

Optimization is performed using flow-matching with the standard mean squared error (MSE) loss and the AIBL reweighted variant-at a constant learning rate of 5×10^{-5}. The training is executed in mixed precision to optimize memory efficiency and converges after 3,000 iterations.

During inference, we employ Gemini 3 Flash as a layout designer to infer spatial bounding boxes from the input prompt. These layouts are then passed to the diffusion transformer for layout-conditioned generation.

C Additional Qualitative Results

In Figure 10, we provide further qualitative comparisons against state-of-the-art models (SDXL, FLUX.1-dev, Z-Image) and our unconditioned baseline across a diverse set of counting prompts. Consistent with our main results, our proposed method reliably separates overlapping objects and adheres closely to requested counts whereas competing models frequently exhibit severe instance fusion or omit objects entirely.

D Qualitative Ablation Study

To illustrate the effect of our proposed components, we provide visual comparisons in Figure 11 for five dense prompts across three configurations: the unconditioned **Baseline** (FLUX.2), **Layout Attn.** (standard layout attention), and our **Layout Attn. + AIBL** model.

The comparisons highlight the specific failure modes of prior approaches and how AIBL mitigates them:

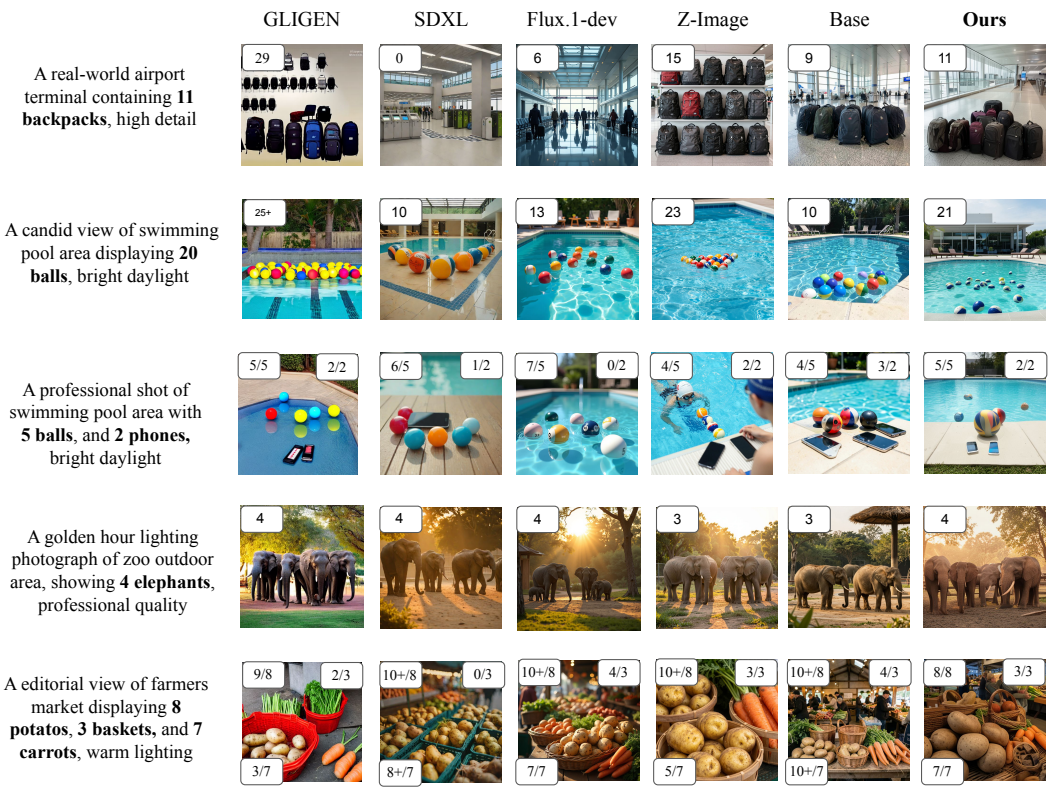


Figure 10: Additional qualitative results comparing our approach against SDXL, FLUX.1-dev, Z-Image, and the unconditioned Baseline on challenging, dense prompts.

- **Baseline.** The baseline model frequently generates incorrect quantities or ignores numerical constraints entirely. For instance, it generates 3 bowls instead of 2 (Row 1), misses an armchair and a plant (Row 2), drops the requested ball (Row 3), and creates an endless, uncounted row of cereal boxes and cans (Row 5).
- **Layout Attn.** Layout attention anchors objects to spatial regions but suffers from severe instance fusion in dense environments. It merges bottles and bowls into indistinct shapes (Row 1), generates an extra sofa-like structure (Row 2), fuses teddy bears and cars into a cluttered mass (Row 3), and blends coffee mugs together (Row 4).
- **Layout Attn. + AIBL (Ours).** Our method resolves these densely overlapping configurations. By rebalancing learning gradients for occluded regions, the model preserves instance separation and produces clear object boundaries. As a result, it correctly renders the requested counts of bottles, armchairs, toy cars, coffee mugs, and cereal boxes without semantic bleeding.

Soft Guidance and Prompt-Layout Interaction. Our layout attention mechanism acts as a *soft spatial prior* rather than a strict bounding-box constraint. The layout suggests where objects should appear, while the text prompt determines what objects to generate and how many.

This is evident in Row 3 of the figure. While the layout incorrectly assigns two regions for "balls" (the prompt specifies only "1 ball"), our model still correctly generates exactly one. In this case, text-conditioned counting overrides conflicting layout cues, allowing the model to produce coherent scenes even when the layout is imperfect. Such layout errors are rare in practice, as layout generation is relatively straightforward and Gemini generally produces reliable layouts for this task.

Limitations. Despite the overall improvements, certain failure cases persist. In Row 5, the model generates only 1 bottle instead of the requested 2. In cluttered multi-category scenes - where dominant object types (e.g., 6 cereal boxes) occupy most of the spatial budget - minor categories can still be under-counted. Addressing per-category attention scaling to ensure balanced generation across all object types remains an open direction for future work.

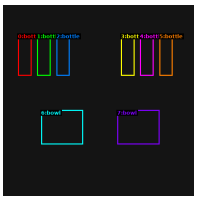
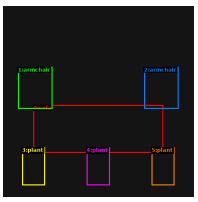
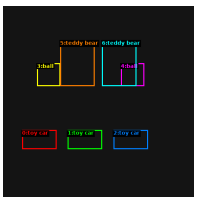
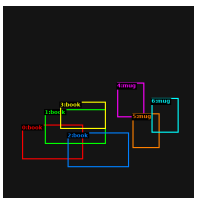
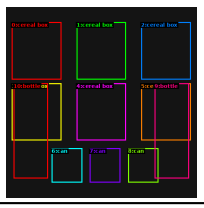
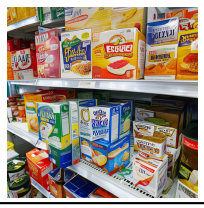
Prompt	Layout	Baseline	Layout Attn.	Layout Attn. + AIBL
*"A kitchen counter with **6 bottles** and **2 bowls**, natural lighting."*				
*"A modern living room with **1 large sofa**, **2 armchairs**, and **3 potted plants**, warm indoor lighting, photo-realistic."*				
*"A clean studio shot of **3 toy cars**, **1 ball**, and **3 teddy bears** on a bright floor."*				
*"A cozy desk scene with **4 books** and **3 coffee mugs**, soft indoor lighting."*				
*"A grocery shelf with **6 cereal boxes**, **3 cans**, and **2 bottles** neatly arranged, sharp focus, realistic photo."*				

Figure 11: **Qualitative Ablation Comparison.** The unconditioned baseline fails to accurately enumerate or arrange objects specified in the prompt. Applying spatial layout attention anchors objects to regions but frequently causes instance fusion in dense areas. Our proposed method (Layout Attn. + AIBL) rebalances gradients for occluded regions, preserving instance separation and producing clearly bounded objects even under heavy occlusion.

E Human Evaluation of Automated Metrics

To validate the reliability of our automated SAM3-based evaluation pipeline, we conducted a manual human evaluation on a random subset of 100 generated samples from our test set. Evaluators were asked to independently count the generated objects and identify cases of object merging or hallucination. We then compared these human-annotated counts against the automated predictions provided by our zero-shot detection pipeline.

As shown in Table 5, the divergence between human evaluation and SAM3 detection is minimal, confirming that our automated metrics provide a rigorous and accurate assessment of counting performance even in densely populated scenes.

To further illustrate the alignment between the target prompts, generated results, and the automated evaluation, Figure 12 provides visual examples of the SAM3 predictions overlaid on our model's

Table 5: Comparison between human evaluation and our automated SAM3 zero-shot detection pipeline on a subset of 100 generated samples. Results indicate close alignment between manual and automated counts.

Evaluation Method	**Exact Match (%)** ↑	**MAE** ↓	**RMSE** ↓
Automated (SAM3)	13.00	4.62	8.75
Human Evaluators	14.00	4.80	9.12
Divergence	1.00	0.18	0.37

generated images. The detector successfully isolates individual instances even under severe occlusion, closely matching human perception.



"An active basketball court featuring 19 bottles, and 16 balls..."
Tar: (19 bottles, 16 balls) | **SAM3**: (25 bottles, 18 balls) | **Hum**: (24 bottles, 15 balls)

"A wide-angle view of home bathroom displaying 3 mirrors..."
Tar: (3 mirrors, 5 plants) | **SAM3**: (3 mirrors, 9 plants) | **Hum**: (3 mirrors, 9 plants)

"A bright daylight photograph of clothing boutique..."
Tar: (4 shoes, 7 backpacks) | **SAM3**: (8 shoes, 6 backpacks) | **Hum**: (6 shoes, 5 backpacks)

"A real-world modern living room containing 15 plants..."
Tar: (15 plants) | **SAM3**: (11 plants) | **Hum**: (9 plants)

"A warm lighting photograph of home office, showing 10 lamps..."
Tar: (10 lamps) | **SAM3**: (10 lamps) | **Hum**: (10 lamps)

"An authentic family dining room scene with 1 table, and 11 candles..."
Tar: (1 table, 11 candles) | **SAM3**: (1 table, 10 candles) | **Hum**: (1 table, 10 candles)

Figure 12: Visual examples of SAM3 predictions overlaid on our generated images. Each sample includes the original generation prompt alongside the target object count, the automated SAM3 detection count, and the manual human evaluation count.

F Additional Experimental Details

F.1 GLIGEN-Style Baseline

We adapt GLIGEN-style grounding [25] to FLUX.2-Klein-base-4B by adding a text/box projector and gated self-attention/dense residual branch after each of its five dual-stream blocks. Unlike the original GLIGEN baseline in Table 1, this reimplementation uses the same backbone and experimental settings as ours. Our method reduces RMSE from 7.80 to 5.99 (23.2%; Table 6).

Table 6: Same-backbone comparison on the 1,000-image OverlapDepth-45K test set. GLIGEN-style denotes our reimplementation. Accuracies are percentages.

Method	Exact ↑	$\pm 10\%$ ↑	$\pm 20\%$ ↑	MAE ↓	RMSE ↓
GLIGEN-style (our reimplementation)	14.20	19.00	31.80	4.55	7.80
Ours	**15.32**	**20.12**	**33.11**	**3.89**	**5.99**

F.2 Hyperparameters and Evaluation

We select the attention strengths using a separate 50-image development set. As shown in Table 7, $(\omega_{\text{intra}}, \omega_{\text{inter}}, \beta) = (2, -1, 2)$ achieves the lowest MAE and RMSE. These values are fixed for all subsequent experiments.

Table 7: Attention-strength sweep on the 50-image development set.

ω_{intra}	ω_{inter}	β	MAE ↓	RMSE ↓
1.0	-1.0	1.5	3.99	8.67
1.0	-1.0	2.0	4.02	8.98
1.0	-2.0	1.5	4.01	8.74
1.0	-2.0	2.0	4.02	8.96
2.0	-1.0	1.5	3.72	7.94
2.0	**-1.0**	**2.0**	**3.65**	**6.41**
2.0	-2.0	1.5	3.76	7.73
2.0	-2.0	2.0	3.96	6.95
3.0	-1.0	1.5	3.84	6.52
3.0	-1.0	2.0	4.06	7.13
3.0	-2.0	1.5	4.13	7.26
3.0	-2.0	2.0	4.25	7.28

G Image-Fidelity Evaluation Protocol

We evaluate 1,000 fixed samples from COCO-3K using identical prompts, seeds, and sampling settings. Our model additionally uses Gemini 3 Flash layouts. Table 8 reports image-fidelity metrics. KID variability is estimated from 100 random subsets of 100 images. Our method improves CLIPScore [13], ImageReward [36], and MUSIQ [18], with comparable KID.

Table 8: Image-fidelity evaluation on COCO-3K. KID values are scaled by 10^3. Subset KID reports mean $\pm$ SD; lower is better.

Model	KID (1K)↓	KID (Subset)↓	CLIPScore↑	ImageReward↑	MUSIQ↑
Pretrained FLUX.2-Klein-4B	**9.037**	**9.430 ± 3.846**	28.424	1.115	73.070
Ours	9.188	9.559 ± 3.125	**28.507**	**1.211**	**74.771**

H Fine-Grained AIBL Ablations

We evaluate two AIBL design choices on a fixed 500-image subset of OverlapDepth-45K, keeping the full layout-attention system unchanged. We compare instance-specific weighting against a common foreground weight and per-image normalization against dataset-global normalization. Table 9 reports term-level Exact (per-category count accuracy), image-level Exact (all category counts correct), and absolute count tolerances (±1, ±2), which differ from the relative tolerances used in the main paper.

Table 9: Fine-grained AIBL ablations on the reported fixed subset. Accuracy columns are percentages. Lower is better for MAE and RMSE.

Variant	Term Exact	Image Exact	±1	±2	MAE	RMSE
Full AIBL	**16.77**	**5.90**	**41.59**	**59.68**	**3.987**	7.481
Common foreground weight	16.65	5.30	38.16	53.43	4.337	7.367
Dataset-global normalization	15.38	**5.90**	37.68	53.31	4.282	**7.233**

Full AIBL achieves the highest term-level Exact and tolerance accuracies and the lowest MAE, while tying the best image-level Exact. However, both alternatives achieve lower RMSE, indicating a trade-off between average counting accuracy and large-error sensitivity. No paired significance test is reported.

I Expanded Human-Preference Evaluation

We conduct a separate pairwise human-preference study with 23 evaluators, 100 evaluation samples, and 1,211 judgments. Evaluators compare randomly selected pairs of model outputs based on overall perceptual quality, with ties permitted. Section I reports the Elo ratings and win–loss–tie statistics.

Table 10: Pairwise human-preference evaluation. Elo ratings and win - loss - tie statistics across evaluated methods.

Model	Reported Elo	Reported W–L–T	Score rate (%)
Ours	1140.23	221–130–18	62.33
3DIS	1015.09	160–121–19	56.50
GLIGEN	1011.40	200–212–27	48.63
CreativeLayout	902.85	168–218–21	43.86
MIGC	899.31	132–304–35	31.74

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The main claims in the abstract and introduction regarding the Ownership-Aware Layout Attention, Amodal-Aware Instance-Balanced Loss (AIBL), and the introduction of the OverlapDepth-45K dataset are directly validated by the quantitative and qualitative results in the experiments section.
   Guidelines:
   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: The paper includes a dedicated limitation section that addresses the model's struggles with density saturation, category-dependent difficulty (e.g., failing to generate thin objects like umbrellas), and synthetic-to-real domain shifts.
   Guidelines:
   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [N/A]
   Justification: This paper introduces an empirical architecture modification and a loss reweighting scheme; it does not present any formal theoretical proofs or theorems.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The authors specify the details necessary to reproduce the experiments, including the automated strict evaluation protocol leveraging zero-shot detection pipelines (SAM 3) and specific hyperparameters like LoRA rank, learning rate, and training steps.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will publicly release the dataset and the source code after the review process.

   Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?
   Answer: [Yes]
   Justification: The paper clearly outlines the dataset splits (44,000 training images, 1,000 test images) and optimization hyperparameters for training the proposed method.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.
7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [No]
   Justification: The paper omits statistical significance reporting due to long execution times of computationally heavy diffusion frameworks.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.
8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
Answer: [Yes]
Justification: Experiments are conducted on a single NVIDIA A100 GPU.
Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**
Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?
Answer: [Yes]
Justification: The paper involves synthetic data generation and does not violate ethical guidelines.
Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**
Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
Answer: [Yes]
Justification: The negative societal risks of the work are discussed in the Appendix, specifically the potential for the framework to be misused to create deceptive content or for malicious profiling and discrimination. It also outlines mitigation strategies, such as advocating for digital watermarking techniques like SynthID to ensure image provenance, and establishes the positive intended use of the framework as strictly for creative and educational purposes.
Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks,

mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The research does not release high-risk models or scraped private data that require specific safeguards against misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly cites all existing models and datasets used, including MS-COCO, T2I-Compbench, SAM 3, and FLUX, etc.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The newly introduced OverlapDepth-45K dataset is well-documented, and details the "Realistic-then-Relabel" generation pipeline, total image counts, and object density statistics.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**
    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
    Answer: [Yes]
    Justification: The paper used human evaluation for the T2I-Compbench dataset.
    Guidelines:
    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.
15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
    Answer: [N/A]
    Justification: The human evaluation involves evaluating generated images for object counts, which does not pose a physical or psychological risk to participants and generally does not require an IRB approval.
    Guidelines:
    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.
16. **Declaration of LLM usage**
    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.
    Answer: [Yes]
    Justification: The paper uses an "LLM layout designer" during inference to generate the input bounding boxes.
    Guidelines:
    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.