CROGEN: PROBING SPATIAL CONTROL AT EXTREME TARGET DENSITIES IN IMAGE GENERATION

Anonymous authors
Paper under double-blind review



Figure 1: We present CroGen, a simple and elegant method for dense spatial control. As the input layouts become increasingly dense, our method keeps generated instances aligned with their target locations, while spatial consistency deteriorates in OminiControl.

ABSTRACT

Advanced image diffusion models generate realistic images and support flexible spatial control. However, we find that spatial correspondence deteriorates as the number of instances we aim to control increases, despite visually convincing outputs. We term this phenomenon spatial constraint overload. In this paper, we explore this phenomenon using dense crowds as a case study and introduce CroGen, a simple and elegant method to alleviate it. Specifically, CroGen combines Proximity Field Conditioning (PFC) and Pixel-Packing Injection (PPI). Our PFC converts target locations into a quantized distance field, providing dense guidance while keeping individual locations explicit. Our PPI retains fine spatial details through pixel packing and injects the packed condition through a single linear projection. We further introduce a dataset of 71K crowd images spanning diverse crowd densities, with point annotations and text prompts. Comprehensive experiments on controllable generation and downstream dense crowd counting show that our CroGen improves spatial consistency without sacrificing image quality. In particular, our method better maintains this consistency as the number of controlled instances grows, demonstrating its effectiveness in alleviating spatial constraint overload. The dataset and model will be released.

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

1 INTRODUCTION

Diffusion models, especially image diffusion models, produce strikingly realistic images with rich visual detail (Ho et al., 2020; Rombach et al., 2022; Esser et al., 2024; Xie et al., 2025). They also demonstrate promising instruction-following capabilities, generating scenes that reflect the object placement and spatial relationships specified through text alone (Saharia et al., 2022).

To enable more flexible control over image generation, existing methods complement text prompts with explicit spatial conditions. ControlNet and T2I-Adapter inject features from structural maps, such as edges and segmentation masks, into a pretrained generator (Zhang et al., 2023; Mou et al., 2024). Instance-level methods, including GLIGEN and InstanceDiffusion, bind object descriptions to specified locations (Li et al., 2023; Wang et al., 2024). Training-free approaches such as BoxDiff guide sampling by imposing spatial constraints on attention maps (Xie et al., 2023).

Despite remarkable progress in controllable generation, we find that spatial consistency with the input conditions deteriorates as the number of instances we aim to control exceeds a certain range, as shown in Fig. 1. Although the generated images still look realistic, objects increasingly appear at incorrect locations, merge with nearby instances, or fail to appear altogether. This reveals a gap between generating convincing dense scenes and maintaining spatial control over the vast number of instances within them. We term this loss of spatial correspondence *spatial constraint overload*.

In this paper, we use dense crowds as a case study to explore spatial constraint overload and introduce CroGen, a simple and elegant method to alleviate it. Our method combines Proximity Field Conditioning (PFC) with Pixel-Packing Injection (PPI) to retain fine spatial cues during condition representation and injection. Specifically, our PFC converts target locations into a distance field whose values indicate proximity to the nearest target. Finer quantization near targets preserves detailed position and spacing cues, while coarser intervals describe more distant regions. The resulting condition extends spatial guidance around each target while keeping its location explicit.

In addition, downsampling in existing condition encoders (Zhang et al., 2023; Tan et al., 2025) can further discard fine spatial details, especially when several targets fall within the region covered by a single image token. We therefore develop PPI, a near-lossless condition injection mechanism without bells and whistles. Through an invertible spatial-to-channel rearrangement, PPI packs the pixels of each region into channels. This aligns the condition with the generator's token grid while retaining every value and its relative position. A simple linear projection then adds the packed condition to the image embeddings, providing detailed cues for controlling densely packed instances.

To support training and evaluation, we introduce a dataset of 71K crowd images with point annotations and text prompts. It comprises 68,836 training images and 2,341 validation images across a broad range of crowd densities. Among all images, 41.2% contain at least 100 people, and 15.9% contain at least 300. We use Qwen3.5 (Qwen Team, 2026) to generate a text prompt for each image. The head annotations specify target locations, while the prompts describe the surrounding scene.

We evaluate our CroGen on controllable generation and downstream dense crowd counting, and further assess spatial consistency at different target counts. Our method consistently achieves the best results across all metrics, surpassing the second-best method by 3.2 in Precision and 2.4 in F1. Meanwhile, it achieves 47.7 FID and 0.779 CLIP-I, showing that improved spatial control comes with maintained image quality. In the densest range of 700–1,000 targets, our method exceeds OminiControl by 6.2 in Precision and 3.2 in F1, demonstrating its effectiveness in alleviating spatial constraint overload. In addition, downstream counting provides complementary validation, with MAE and MSE reduced by 8.5 and 11.8 relative to their respective second-best results.

The main contributions of this work are as follows:

- We reveal spatial constraint overload, where spatial correspondence deteriorates as the number of instances we aim to control increases, despite visually realistic outputs. We explore this phenomenon using dense crowds as a case study.
- We propose CroGen, a simple and elegant framework that combines Proximity Field Conditioning and Pixel-Packing Injection. Our PFC provides dense spatial guidance around target locations, while PPI retains condition details for direct injection.

- We introduce a dataset of 71K crowd images across diverse densities, with point annotations and text prompts. Generation and downstream counting experiments show that our CroGen alleviates spatial constraint overload more effectively than prior methods.

2 Related Work

2.1 Spatially Controllable Image Generation

Modern image generation models have demonstrated strong controllability, enabling users to steer generation beyond semantic descriptions with explicit structural and visual conditions. As an early representative approach, ControlNet (Zhang et al., 2023) augments a frozen diffusion model with trainable conditioning branches and zero-initialized connections, enabling spatial control through signals such as edges, depth, and human poses. T2I-Adapter (Mou et al., 2024) instead introduces lightweight adapters that align external conditions with the internal representations of a frozen generator, providing parameter-efficient control over structures and colors. Together, they establish practical mechanisms for incorporating explicit spatial guidance into pretrained image generators.

More recent methods extend controllable generation to Diffusion Transformers (DiTs) (Peebles & Xie, 2023), with increasing emphasis on unified and efficient conditioning mechanisms. Omini-Control (Tan et al., 2025) reuses the pretrained VAE and DiT blocks to encode conditional images and jointly process condition and image tokens. Its dynamic positional encoding supports both spatially aligned and non-aligned conditions with minimal additional parameters. EasyControl (Zhang et al., 2025) introduces lightweight LoRA modules to incorporate external conditions into DiTs. Combined with position-aware training and efficient attention, it supports flexible resolutions, plug-and-play control, and zero-shot composition of multiple conditions. Recent progress has broadened the scope of controllable generation, enabling increasingly diverse conditions to be integrated in a unified and efficient manner. However, dense instance-level constraints remain a distinct challenge.

2.2 Instance-Level and Dense Spatial Control

Instance-level generation provides finer spatial control by associating individual objects with explicit locations. GLIGEN (Li et al., 2023) injects grounded object descriptions and bounding boxes into a frozen diffusion model through gated layers. BoxDiff (Xie et al., 2023) achieves box-level control without additional training by constraining cross-attention maps during sampling. InstanceDiffusion (Wang et al., 2024) further supports diverse location specifications, including points, scribbles, boxes, and masks. For denser and more occluded layouts, Ngo et al. (Ngo et al., 2026) introduce layout-aware attention biases and an amodal-balanced loss to mitigate cross-instance interference and weak supervision for occluded objects. This line of study progressively extends spatial control to complex multi-instance configurations.

As the number of instances grows, related studies have also explored quantity control and dense spatial conditioning. Counting Guidance (Kang et al., 2025) uses gradients from an external counting network to steer diffusion sampling toward a requested object count, while CountGen (Binyamin et al., 2025) exploits instance-aware features within diffusion models to improve count accuracy. YOLO-Count (Zeng et al., 2025) further develops a differentiable open-vocabulary counter that provides gradient-based quantity guidance for image generation. More closely related, diffusion-based data augmentation for object counting conditions ControlNet on smoothed point-density maps, supplemented by a counting objective and guided sampling to improve location consistency (Wang et al., 2025). DenseControl (Wang et al., 2026) explicitly targets dense instance layouts through isolated object embeddings, scale encoding, and positional enhancement. Yet how spatial correspondence deteriorates as target locations become increasingly dense remains largely unexplored. We study this failure mode as spatial constraint overload and show that preserving fine position cues through both condition representation and injection helps alleviate it.

3 Method

To alleviate spatial constraint overload, we present CroGen, a simple framework that preserves fine spatial information throughout condition representation and injection. As illustrated in Fig. 3, Prox-

(a) Source Dataset Splits

Dataset	Train*	Val*	Test
SHA	300	–	182
SHB	400	–	316
JHU++	2,272	500	1,600
NWPU	3,109	500	1,500
UCF-QNRF	1,201	–	334
UCF-CC-50	50	–	0

* Splits used to construct the training set.

(b) Crowd Size Distribution

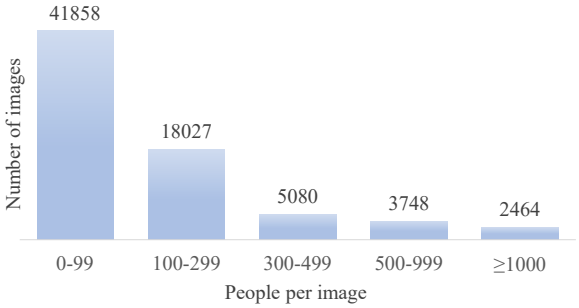


Figure 2: Dataset statistics. (a) Source dataset splits. (b) Crowd-size distribution.

imity Field Conditioning (PFC) extends target locations into dense spatial guidance. Pixel-Packing Injection (PPI) retains the condition values and their spatial arrangement through invertible pixel packing, then adds their linear projections directly to the image embeddings.

3.1 Data Preparation

We collect training images from the training and validation splits of ShanghaiTech (Zhang et al., 2016), JHU-CROWD++ (Sindagi et al., 2020), NWPU-Crowd (Wang et al., 2020), and UCF-QNRF (Idrees et al., 2018). Their test splits form the held-out pool. We use all 50 UCF-CC-50 images for training (Idrees et al., 2013). Fig. 2(a) lists the original split sizes before preprocessing. We retain the original head annotations as spatial conditions.

We create five fixed crops from the four corners and center of each training image, supplemented with random crops. Perceptual hashing removes near-duplicates. The resulting dataset contains 71,177 images, with 68,836 for training and 2,341 for validation. The processed dataset spans a broad range of crowd densities, from fewer than 100 to at least 1,000 people per image. While 58.8% of the images contain fewer than 100 people, 29,319 images (41.2%) contain at least 100, including 6,212 (8.7%) with at least 500. For text conditioning, we use Qwen3.5 (Qwen Team, 2026) to generate a text prompt for each image.

3.2 Proximity Field Conditioning

To provide dense spatial guidance while preserving individual target locations, we propose Proximity Field Conditioning (PFC). Specifically, our PFC converts target locations into a quantized distance field, where each pixel encodes proximity to the nearest target, as shown in Fig. 3. The field extends guidance around each target without summing responses from neighboring instances, keeping individual locations explicit even when targets are densely packed.

Given a nonempty set of target head locations $\mathcal{P}$ on the image grid, we define the distance field as

$$D(u) = \min_{p_i \in \mathcal{P}} \|u - p_i\|_2, \tag{1}$$

where u is a pixel location, p_i is a target location, and $\|\cdot\|_2$ measures Euclidean distance in pixels. Each target appears as a distance minimum, while the surrounding values describe proximity to its nearest target. Without accumulating responses, the field retains a consistent distance interpretation across crowd densities.

To retain more detail around closely spaced targets, our PFC uses finer distance intervals near their locations and coarser intervals farther away. This allocates more distance levels to the regions where small spatial differences distinguish neighboring targets. We quantize the field as

$$C(u) = \sum_{k=1}^{K} \mathbf{1}[D(u) \geq \tau_k], \tag{2}$$

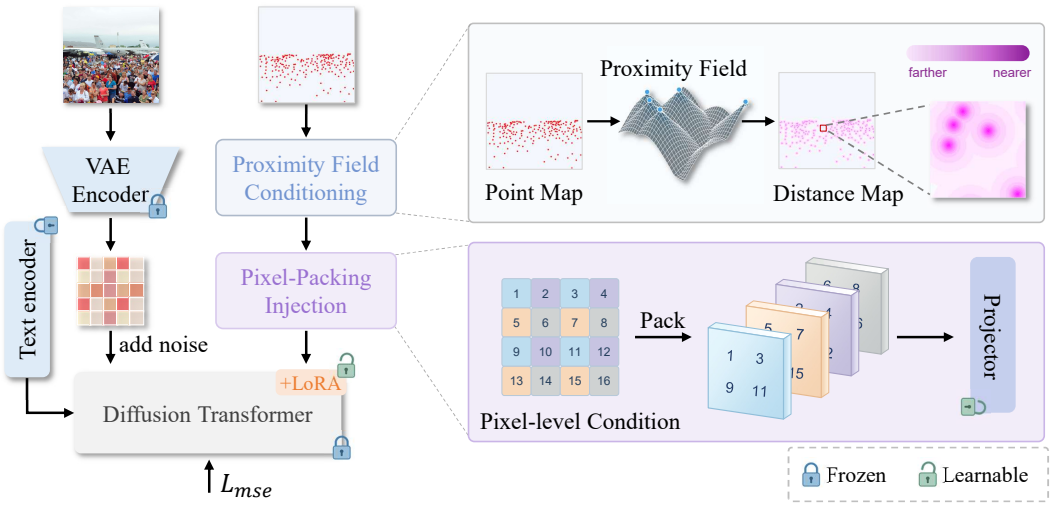


Figure 3: Overview of our CroGen. PFC converts target locations into a quantized distance map, providing dense guidance without accumulating responses from nearby targets. PPI packs map pixels into channels while retaining their values and relative positions.

where K is the number of thresholds, $0 < \tau_1 < \cdots < \tau_K$, and $\mathbf{1}[\cdot]$ equals one when its condition holds and zero otherwise. Each value $C(u) \in \{0, \ldots, K\}$ records a distance level. Their growing intervals retain fine position cues near targets and coarser context elsewhere.

3.3 Pixel-Packing Injection

We further introduce Pixel-Packing Injection (PPI), as shown in Fig. 3, a simple mechanism that preserves the spatial details encoded by our PFC during condition injection. Existing condition encoders often downsample condition maps to match the generator's token grid (Tan et al., 2025; Zhang et al., 2025), which can discard fine spatial cues when several targets fall within a single token region. Pixel packing aligns the grids while retaining every condition value and its relative position.

PPI separates spatial alignment from information compression. For a condition $C \in \mathbb{R}^{H \times W}$ with height H and width W, the generator uses a token grid of height H_g and width W_g. With the integer scale factor $r = H/H_g = W/W_g$, PPI packs each $r \times r$ nonoverlapping region into a vector,

$$Z_c = \mathrm{Pack}_r(C) \in \mathbb{R}^{L \times r^2}, \qquad L = H_g W_g. \tag{3}$$

Here, Pack_r rearranges pixels into channels in a fixed order and flattens the grid in the same order as the image tokens. The resulting Z_c contains L condition tokens, each with r^2 values. Each channel corresponds to a fixed pixel position within its region, so packing preserves every value and is exactly invertible.

This alignment allows direct injection through a single linear projection and additive fusion,

$$E_t = \mathrm{Proj}(Z_t) + Z_c W_c^\top, \tag{4}$$

where Z_t contains the noisy image tokens at timestep t, and $\mathrm{Proj}(\cdot)$ is the learned image projection that maps each token to embedding dimension d. The trainable matrix $W_c \in \mathbb{R}^{d \times r^2}$ projects condition tokens to the same dimension without a bias. Their sum $E_t \in \mathbb{R}^{L \times d}$ forms the conditioned image embeddings supplied to the generator.

Pixel packing retains the spatial details that distinguish nearby targets, even when they share the same image token. A simple linear projection introduces this information directly into the image embeddings, without an intermediate downsampling encoder.

3.4 Learning Objective

We train our CroGen with the standard flow matching objective (Lipman et al., 2023; Liu et al., 2023). Let Z_0 denote the clean image latent produced by the frozen image autoencoder and arranged as tokens. At a sampled timestep t, we interpolate between Z_0 and noise ϵ of the same shape, whose entries are independent standard normal samples,

$$Z_t = (1 - \sigma_t)Z_0 + \sigma_t \epsilon, \tag{5}$$

where $\sigma_t \in [0, 1]$ is the noise level at timestep t. The generator takes the conditioned embeddings E_t, the encoded text prompt s, and t to predict a velocity $\hat{v}_t$ with the same shape as Z_0. Along this linear path, the target velocity with respect to the noise level is $\epsilon - Z_0$. We minimize

$$\mathcal{L}_{\mathrm{FM}} = \mathrm{MSE}(\hat{v}_t, \epsilon - Z_0)\,. \tag{6}$$

Here, MSE averages the squared differences over all token and channel entries. Training averages this loss over image–condition pairs, sampled timesteps, and noise.

4 Experiments

4.1 Evaluation Settings and Metrics

Evaluation Data. For the main spatial consistency and counting comparisons, we construct 5,000 conditions by randomly cropping four 512×512 patches from the training condition maps, combining them into a 2×2 grid, and applying random spatial perturbations. Qwen3.5 (Qwen Team, 2026) randomly generates a text prompt for each condition. Image quality is evaluated separately using the 2,341 validation images as real references and their head locations and text prompts as generation conditions. Within each comparison, methods share target locations, prompts, and sampling seeds.

Spatial Consistency (Spatial Cons.). We use Precision and F1 to measure spatial consistency between generated instances and their target locations. Specifically, a fixed P2PNet (Song et al., 2021) detects heads in the generated images, and we match the detected positions to the conditioning points. We report Precision and F1 with both scores multiplied by 100. Together, they measure correspondence at individual target locations, exposing spatial errors that agreement in the total count can conceal. The main comparison uses one image per condition. We also generate 5,000 images in each of five target count ranges, 0–99, 100–299, 300–499, 500–699, and 700–1,000.

Image Quality. We use Fréchet Inception Distance (FID) (Heusel et al., 2017) and CLIP image similarity (CLIP-I) (Radford et al., 2021) to assess visual quality and agreement with reference scenes. Specifically, FID measures the distance between the feature distributions of generated and real images, while CLIP-I computes the cosine similarity between the CLIP embeddings of each generated image and its reference.

Crowd Counting. We further use generated images in downstream crowd counting to validate the effectiveness of our method. The conditioning points serve as annotations, so better spatial consistency reduces mismatches between generated instances and their labels. This provides more reliable supervision and can improve the trained counter's performance on real images. We train a P2PNet on the 5,000 images generated by each method for the main spatial consistency evaluation. All counters are trained for 100 epochs under the same protocol and evaluated on the complete ShanghaiTech Part A test set (Zhang et al., 2016). We report MAE and MSE following the P2PNet convention.

4.2 Implementation Details

We initialize our CroGen from FLUX.1-dev (Labs, 2024) and train it at a resolution of 1024×1024. We fine-tune its DiT using LoRA (Hu et al., 2022) with rank 64 and jointly optimize the condition projection. We use Adam (Kingma & Ba, 2015) for optimization. The learning rate increases linearly to 10^{-4} over the first 500 iterations and then decays linearly. All generators are trained on our crowd training set for 25,000 iterations with a batch size of 8. FLUX.1 in Tab. 1 serves as our **Baseline** and uses the Diffusers (von Platen et al., 2022) `flux-control` implementation with a point map as the default condition. For ControlNet (Zhang et al., 2023), we use the default FLUX ControlNet implementation and configuration in Diffusers. OminiControl (Tan et al., 2025),

Table 1: Quantitative comparisons of controllable image generation and downstream crowd counting. All methods use the same training split and evaluation protocol.

Method	Reference	Spatial Cons.		Image Quality		Counting	
		Prec. ↑	F1 ↑	FID ↓	CLIP-I ↑	MAE ↓	MSE ↓
FLUX.1 (Labs, 2024)	–	61.7	59.0	<u>48.3</u>	<u>0.761</u>	85.4	<u>141.4</u>
ControlNet (Zhang et al., 2023)	ICCV 23	49.5	39.5	69.8	0.731	<u>82.0</u>	143.6
OminiControl (Tan et al., 2025)	ICCV 25	<u>67.6</u>	<u>63.0</u>	51.0	0.758	163.5	281.4
T2I-Adapter (Mou et al., 2024)	AAAI 24	8.0	4.7	57.7	0.690	426.4	550.6
EasyControl (Zhang et al., 2025)	ICCV 25	12.6	10.9	52.3	0.755	353.0	484.5
Diffusion Aug. (Wang et al., 2025)	ICASSP 25	60.1	61.3	76.5	0.645	292.1	433.5
CroGen (Ours)	–	**70.8**	**65.4**	**47.7**	**0.779**	**73.5**	**129.6**

T2I-Adapter (Mou et al., 2024), EasyControl (Zhang et al., 2025), and Diffusion Aug. (Wang et al., 2025) follow their official implementations and default settings under the shared training setup. All experiments are conducted on a system with 8 NVIDIA A100 GPUs.

4.3 Main Results

Tab. 1 presents quantitative results for controllable image generation and downstream crowd counting. Our CroGen consistently achieves the best results across all six metrics. For controllable image generation, our method achieves 70.8 Precision and 65.4 F1, surpassing the second-best method, OminiControl (Tan et al., 2025), by 3.2 in Precision and 2.4 in F1. The gains in both metrics indicate closer correspondence between generated instances and their target locations. The results suggest that our CroGen reduces spatial errors that can remain hidden in visually convincing images. Relative to FLUX.1, our method improves Precision by 9.1 and F1 by 6.4. The gains over the baseline suggest that preserving detailed position cues helps the generator distinguish nearby targets and place them at their target locations. Meanwhile, our CroGen also obtains 47.7 FID and 0.779 CLIP-I, the best results among the evaluated methods. The two metrics suggest that our method maintains the overall visual characteristics of real images and the scene content of the corresponding references. Our method thus improves spatial control while preserving image quality.

To further validate the effectiveness of our method, we generate 5,000 images with each model using the same spatial conditions and text prompts. We then train a P2PNet counter on each generated set and evaluate all counters on the ShanghaiTech Part A test set. Our CroGen achieves the lowest MAE of 73.5 and MSE of 129.6. It reduces MAE by 8.5 relative to ControlNet and MSE by 11.8 relative to FLUX.1, the respective second-best methods. Here, the conditioning points serve as annotations, so missing or misplaced instances can weaken the connection between visual content and supervision. Together, the spatial and counting results indicate that our CroGen improves consistency between generated instances and their target locations.

Fig. 4 presents qualitative comparisons under dense spatial conditions. Existing methods can generate visually plausible scenes while leaving target points on the background or shifting heads away from their target locations. Our CroGen maintains closer correspondence in the enlarged regions, where neighboring instances are also more clearly separated. The examples show that our method alleviates spatial constraint overload while preserving realistic scene content.

4.4 Ablation Study

We examine condition representation and injection using FLUX.1 in Tab. 1 as the baseline. Unless otherwise specified, all variants follow the same training and evaluation settings described above.

Ablation on Condition Map. In Tab. 2, we investigate how different condition maps affect generation under dense spatial control. We compare a point map, a density map, and the distance map produced by our PFC, using the same VAE encoder for all three. Our distance map substantially outperforms the point map in Precision, improving the score by 6.8 from 61.7 to 68.5,

and achieves the highest F1 of 59.6. It also exceeds the Gaussian density map by 3.5 in both Precision and F1. Gaussian responses overlap around nearby targets, which can blur their position cues in crowded regions. In contrast, our PFC measures distance to the nearest target without adding responses from neighboring instances, keeping individual locations explicit. Our distance map also achieves the best FID and CLIP-I, showing that spatial consistency improves without sacrificing image quality. Overall, the comparison shows that the form of the condition map has a substantial impact on spatial consistency in controllable generation. Our distance map offers a simple way to turn a plain point map into a dense condition with richer spatial cues, improving spatial consistency.

Table 2: Effect of condition representation on the FLUX.1 baseline. All maps use the same pretrained VAE encoder. Our distance map is produced by our PFC.

Condition Map	Spatial Cons.		Image Quality		Counting	
	Prec. ↑	F1 ↑	FID ↓	CLIP-I ↑	MAE ↓	MSE ↓
Point Map	61.7	59.0	48.3	0.761	85.4	**141.4**
Density Map	65.0	56.1	49.4	0.764	83.2	145.3
Distance Map (Ours)	**68.5**	**59.6**	**47.8**	**0.769**	**80.0**	149.1

Ablation on Our Components. In Tab. 3, we evaluate our PFC and PPI by adding them to the baseline in sequence. Our PFC replaces the point map with a distance map and improves Precision from 61.7 to 68.5. Building on our PFC, our PPI injects the distance map directly without VAE encoding of the condition map. Our PPI consistently improves performance across all six metrics, with F1 increasing by 5.8 from 59.6 to 65.4 and MSE decreasing by 19.5 from 149.1 to 129.6. Precision also increases from 68.5 to 70.8. The additional gains show that how spatial information enters the generator matters even when the condition map remains the same. Our PPI packs condition values within each token region into separate channels before projection, preserving detailed position cues during alignment with the token grid. This gives the generator access to spatial distinctions between nearby targets that share a single token region. Dense guidance thus becomes more effective when its spatial details are also preserved during injection, helping our CroGen alleviate spatial constraint overload.

Table 3: Ablation of our PFC and PPI on the FLUX.1 baseline. ✓ indicates an enabled component. The baseline encodes a point map using the pretrained VAE.

PFC	PPI	Spatial Cons.		Image Quality		Counting	
		Prec. ↑	F1 ↑	FID ↓	CLIP-I ↑	MAE ↓	MSE ↓
–	–	61.7	59.0	48.3	0.761	85.4	141.4
✓	–	68.5	59.6	47.8	0.769	80.0	149.1
✓	✓	**70.8**	**65.4**	**47.7**	**0.779**	**73.5**	**129.6**

Comparison across Target Instance Counts. To further explore how image generators perform with condition maps containing different numbers of target instances, we compare results across five target count ranges in Tab. 4, generating 5,000 images per model in each range. Beyond the 0–299 range, both methods show declining Precision and F1 as target counts rise, reflecting the loss of spatial correspondence that characterizes spatial constraint overload. Our CroGen alleviates this deterioration. From 100–299 to 700–1,000 targets, Precision decreases by 9.7 for our method, compared with 15.6 for OminiControl. Our Precision advantage consequently widens from 0.3 to 6.2. In the densest range, our method also outperforms OminiControl by 3.2 in F1. The widening gap

Table 4: Comparison with OminiControl across target instance counts. Each model generates 5,000 images per range.

Range	Method	Spatial Cons.		Counting	
		Prec. ↑	F1 ↑	MAE ↓	MSE ↓
0–99	OminiControl	67.9	62.5	259.3	400.5
	CroGen	**70.9**$_{+3.0}$	**65.7**$_{+3.2}$	**140.6**$_{-118.7}$	**258.7**$_{-141.8}$
100–299	OminiControl	71.0	71.4	110.3	222.4
	CroGen	**71.3**$_{+0.3}$	**72.5**$_{+1.1}$	**75.5**$_{-34.8}$	**138.8**$_{-83.6}$
300–499	OminiControl	67.5	68.3	86.9	162.3
	CroGen	**69.3**$_{+1.8}$	**69.2**$_{+0.9}$	**72.0**$_{-14.9}$	**124.6**$_{-37.7}$
500–699	OminiControl	59.5	58.9	92.4	167.0
	CroGen	**63.1**$_{+3.6}$	**60.7**$_{+1.8}$	**77.8**$_{-14.6}$	**125.0**$_{-42.0}$
700–1,000	OminiControl	55.4	53.6	97.2	189.8
	CroGen	**61.6**$_{+6.2}$	**56.8**$_{+3.2}$	**81.6**$_{-15.6}$	**129.4**$_{-60.4}$



Figure 4: Qualitative comparison under the same spatial conditions. "Ref. Image" denotes reference images. Red points mark target head locations. Our CroGen maintains closer correspondence between generated heads and target locations.

indicates that our method is particularly effective where spatial constraints become dense, and overload is more pronounced. Together with the ablations, this trend suggests that preserving position cues through condition representation and injection helps alleviate spatial constraint overload.

Fig. 1 illustrates how this overload appears in generated images. As control maps become denser, OminiControl produces plausible scenes, yet heads drift from their target points or merge with nearby instances. Our CroGen maintains closer correspondence in the enlarged views, showing why visual realism alone does not guarantee spatial control.

Counting provides complementary evidence because the conditioning points also serve as annotations. Spatial errors can therefore create mismatches between generated images and their labels. Under the same training protocol and real test set, counters trained on images from our CroGen achieve lower MAE and MSE in every target count range. In the densest range, MAE falls from 97.2 to 81.6 and MSE from 189.8 to 129.4 relative to OminiControl. Together with higher Precision and F1, these gains indicate more faithful spatial control.

5 Conclusion

In this paper, we explore spatial constraint overload through dense crowds, where visually convincing images can lose correspondence to target locations as the number of controlled instances grows. To alleviate this deterioration, we introduce CroGen, a simple framework that combines Proximity Field Conditioning (PFC) and Pixel-Packing Injection (PPI). PFC turns target locations into a quantized distance field, extending guidance around each target while retaining the cues that distinguish nearby instances. PPI carries these cues into the generator by packing spatial values into channels aligned with its token grid. We also introduce a dataset of 71K crowd images spanning diverse densities, with head locations and text prompts, to support training and evaluation. Experiments on controllable generation and downstream counting show that our CroGen improves spatial consistency over prior methods, particularly at high target counts, while maintaining image quality.

Limitations. Despite the encouraging progress, we observe that generated instances become highly similar in appearance when multiple target locations fall within a single token region. Our CroGen preserves their position cues but does not explicitly control individual appearance. Joint control of instance positions and appearance is also a vital direction for improving diversity in dense scenes.

AI USE STATEMENT

Generative AI tools were used only for linguistic polishing of the manuscript. They were not used for code generation or modification, research ideas, model or method design, experimental design, analyses, claims, or interpretations. The researchers authored, validated, and cross-checked all algorithms, equations, results, and conclusions. All AI-assisted text was manually reviewed and validated by the authors, who take full responsibility for the final content of this work.

REFERENCES

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 *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proc. of Intl. Conf. on Machine Learning*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proc. of Advances in Neural Information Processing Systems*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. of Advances in Neural Information Processing Systems*, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proc. of Intl. Conf. on Learning Representations*, 2022.

Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2013.

Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proc. of European Conference on Computer Vision*, 2018.

Wonjun Kang, Kevin Galim, Hyung Il Koo, and Nam Ik Cho. Counting guidance for high fidelity text-to-image synthesis. In *Proc. of IEEE Winter Conf. on Applications of Computer Vision*, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of Intl. Conf. on Learning Representations*, 2015.

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

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 *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *Proc. of Intl. Conf. on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proc. of Intl. Conf. on Learning Representations*, 2023.

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 *Proc. of the AAAI Conf. on Artificial Intelligence*, 2024.

Bach-Hoang Ngo, Si-Tri Ngo, Hieu Le, and Trung-Nghia Le. Learning to generate multiple objects from dense and occluded layouts. *arXiv preprint arXiv:2607.03488*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2023.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of Intl. Conf. on Machine Learning*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Proc. of Advances in Neural Information Processing Systems*, 2022.

Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2021.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2025.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Juncheng Wang, Lei Shang, Wang Lu, Baigui Sun, and Shujun Wang. Densecontrol: Instance-level controllable synthesis of dense crowd image. *IEEE Transactions on Multimedia*, 2026.

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

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2024.

Zhen Wang, Yuelei Li, Jia Wan, and Nuno Vasconcelos. Diffusion-based data augmentation for object counting problems. In *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing*, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *Proc. of Intl. Conf. on Learning Representations*, 2025.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2023.

Guanning Zeng, Xiang Zhang, Zirui Wang, Haiyang Xu, Zeyuan Chen, Bingnan Li, and Zhuowen Tu. Yolo-count: Differentiable object counting for text-to-image generation. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2023.

Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2016.

Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2025.