



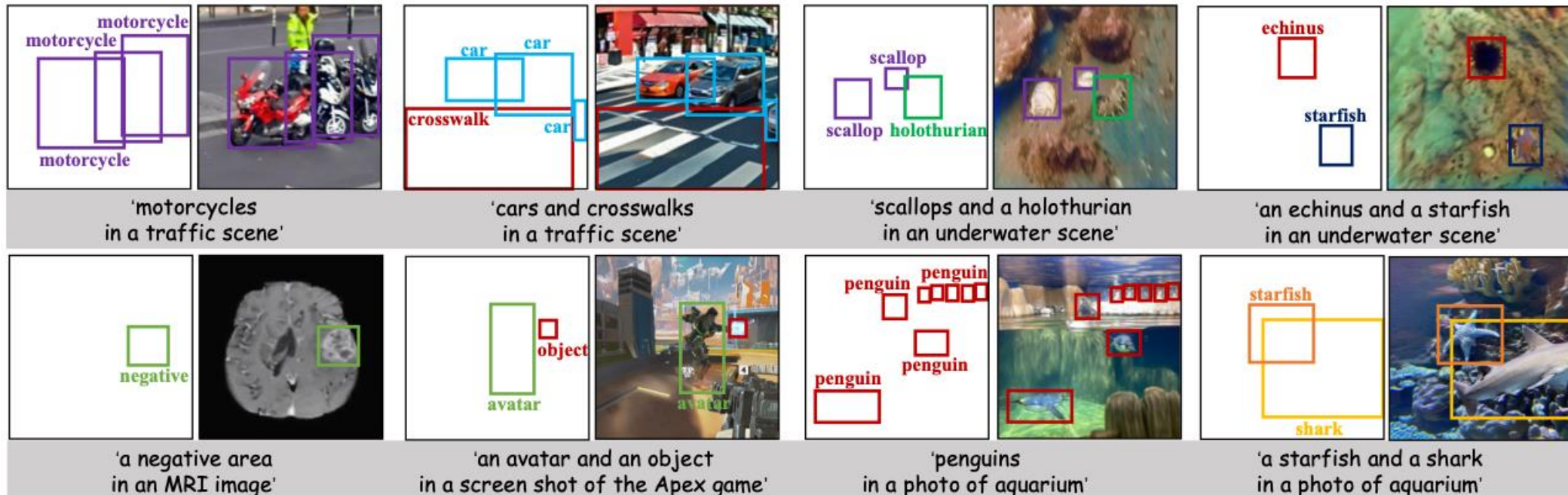
ODGEN: Domain-specific Object Detection Data Generation with Diffusion Models

Jingyuan Zhu, Shiyu Li, Yuxuan Liu, Jian Yuan,
Ping Huang, Jiulong Shan, Huimin Ma



Motivations

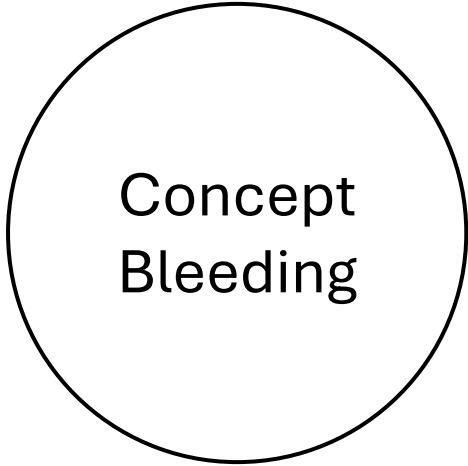
- Generate domain-specific data with diffusion models
- Control generated content with bounding boxes and labels
- Boost object detectors with synthetic images



Challenges



Domain
Gap

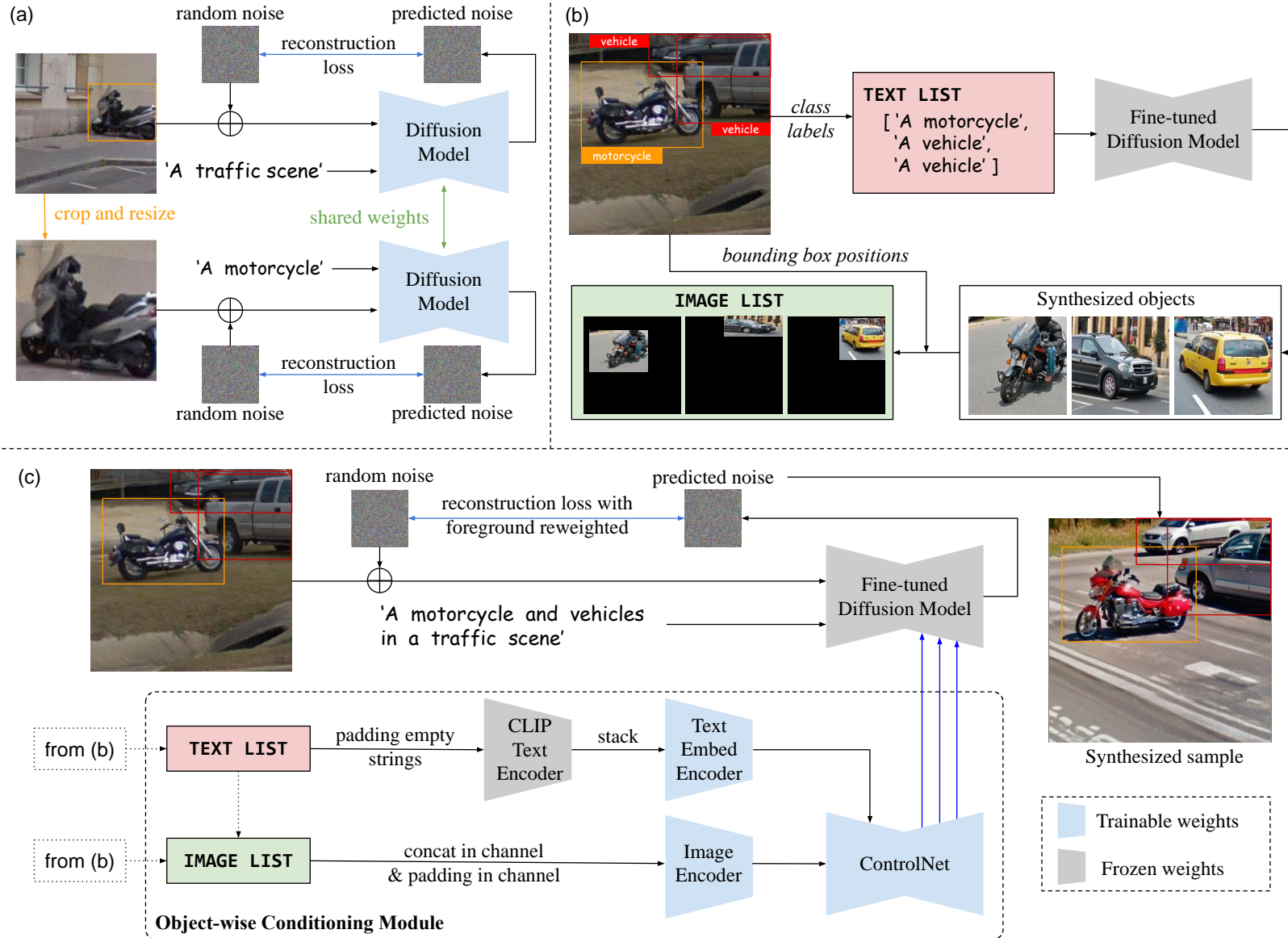


Concept
Bleeding

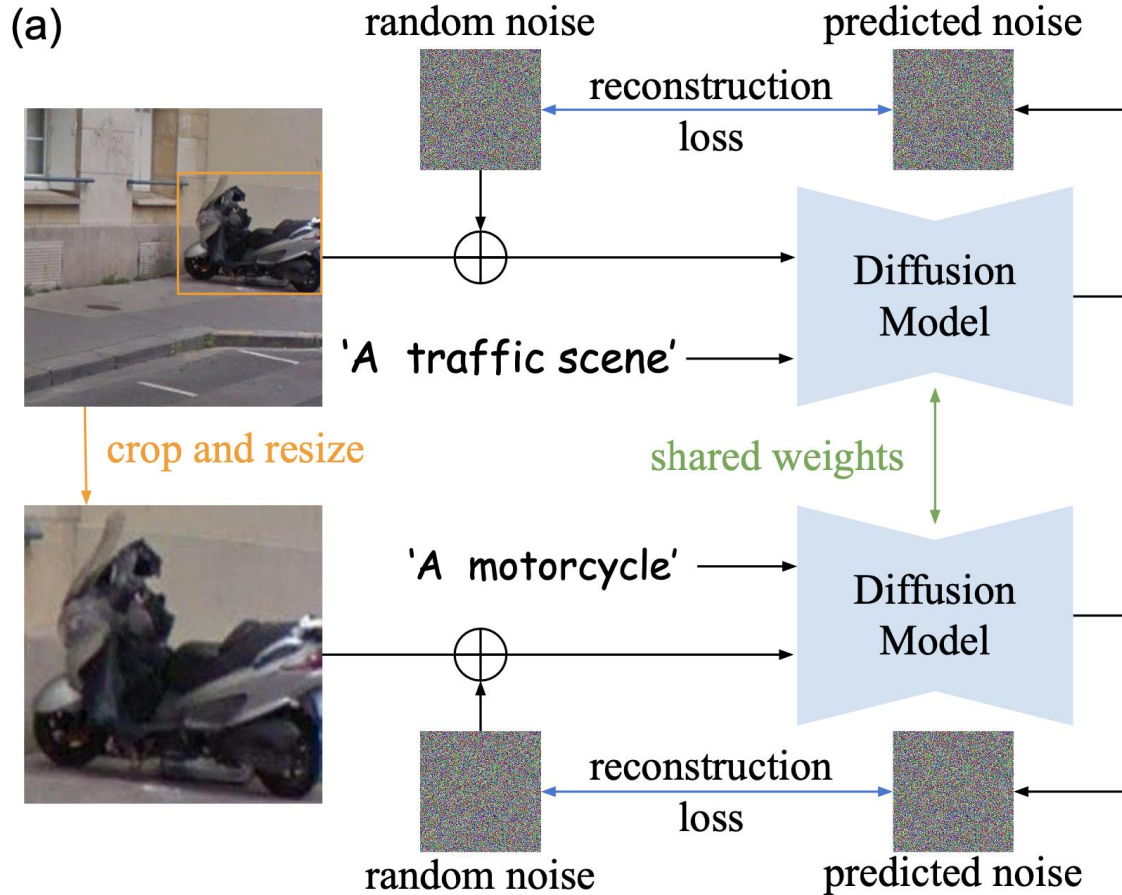


Object
Occlusion

Method Overview

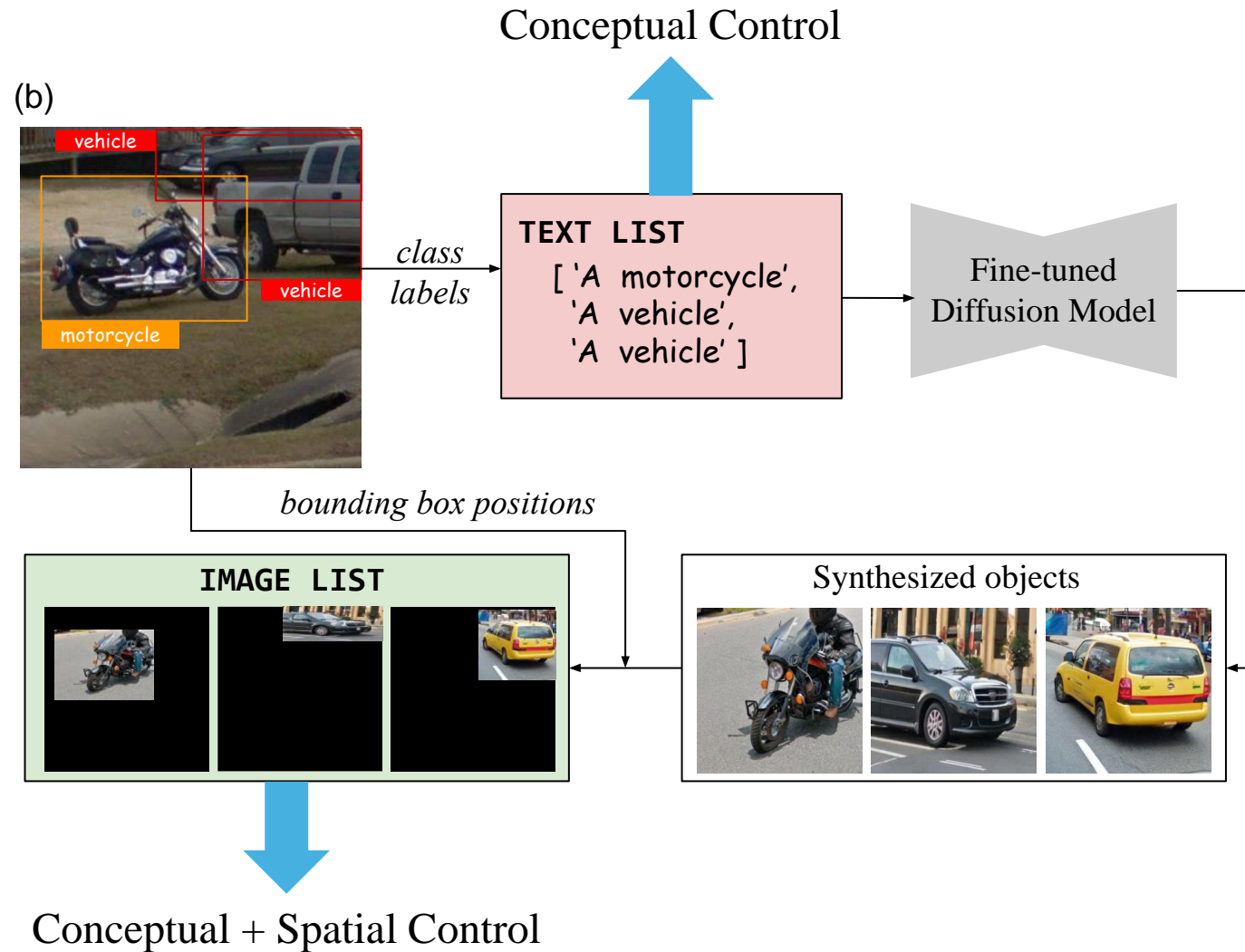


Domain-specific Fine-tuning

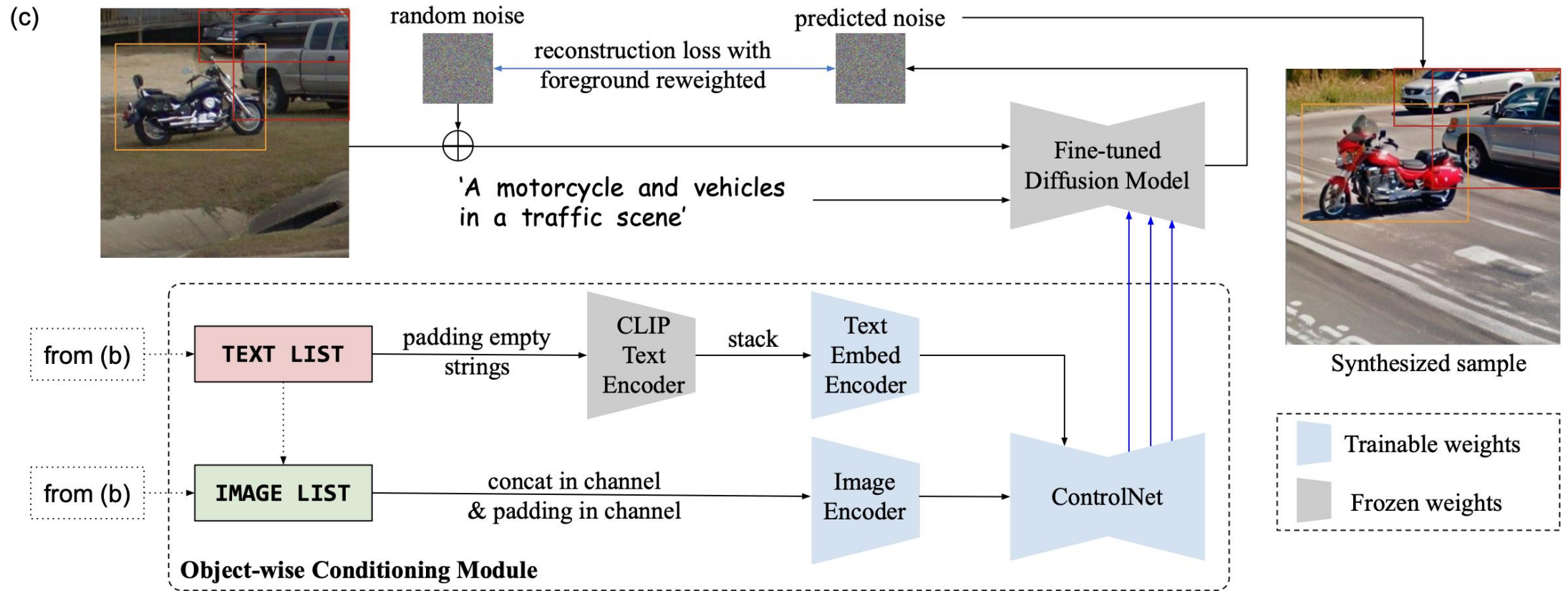


$$\mathcal{L}_{rec} = \mathbb{E}_{x_o, t, \epsilon_o \sim \mathcal{N}(0,1)} [\|\epsilon_o - \epsilon_\theta(x_o^t, t, \tau(c_o))\|^2] \\ + \lambda \mathbb{E}_{x_s, t, \epsilon_s \sim \mathcal{N}(0,1)} [\|\epsilon_s - \epsilon_\theta(x_s^t, t, \tau(c_s))\|^2]$$

Text List & Image List



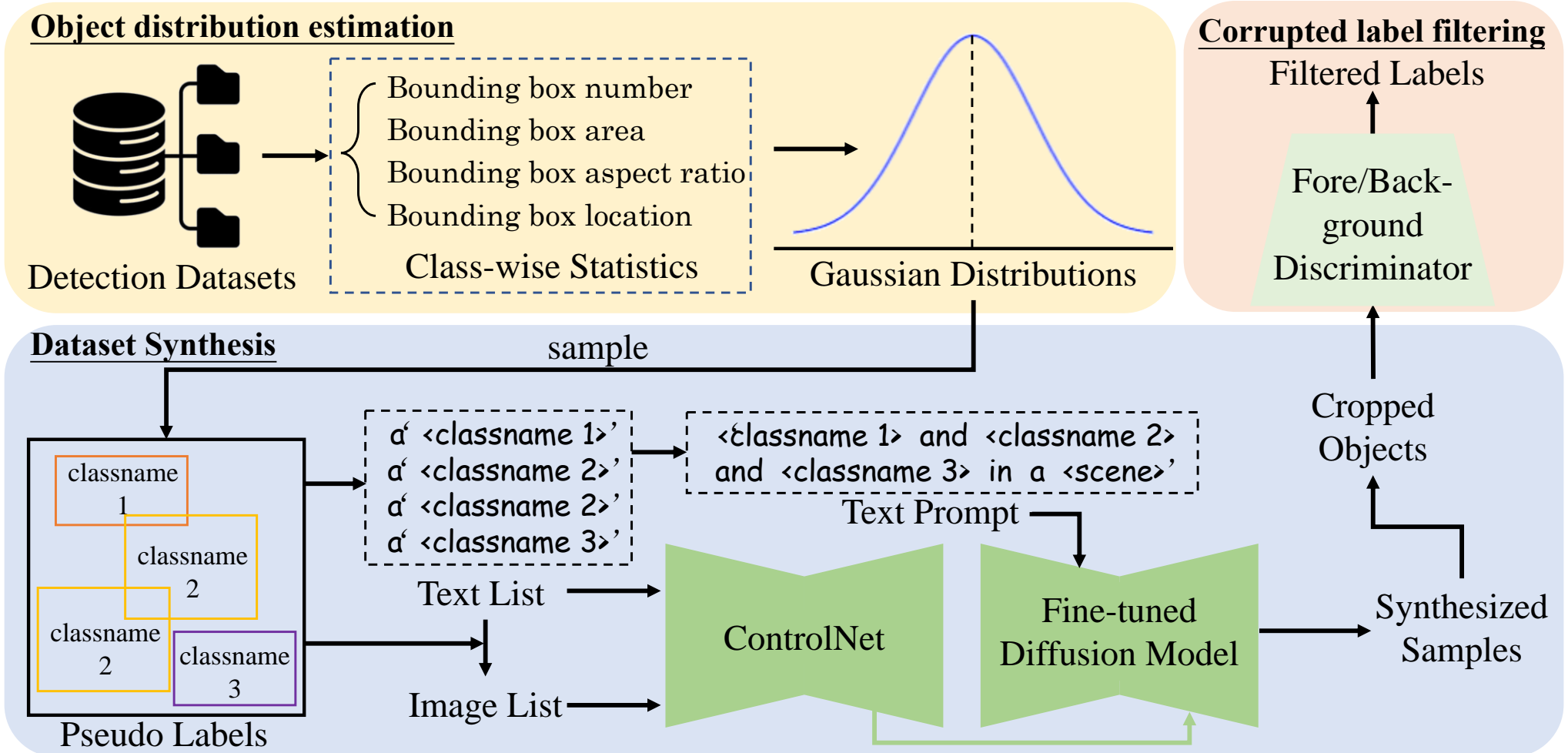
Object-wise Conditioning



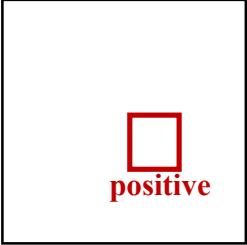
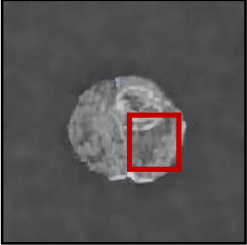
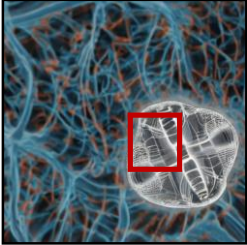
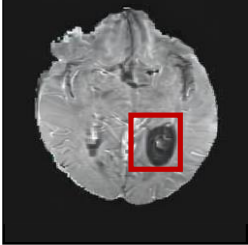

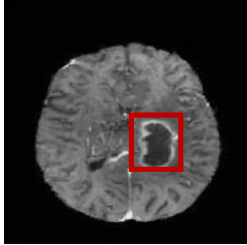
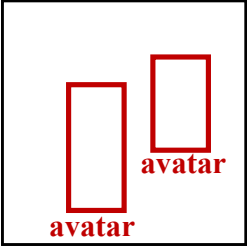





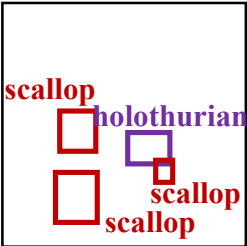
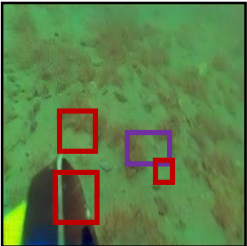
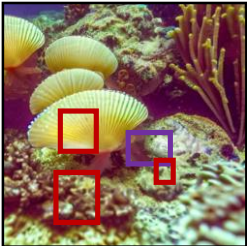
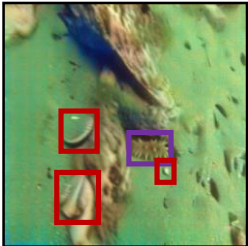
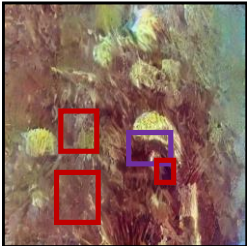
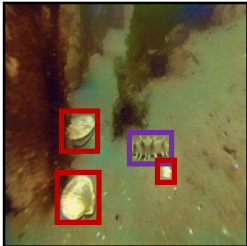
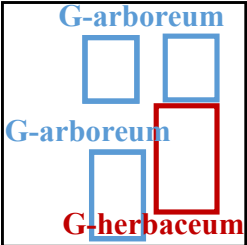
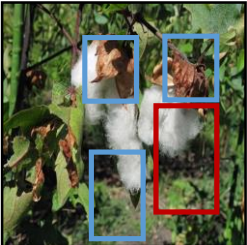
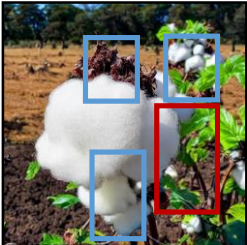
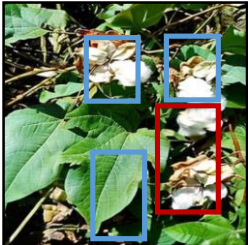
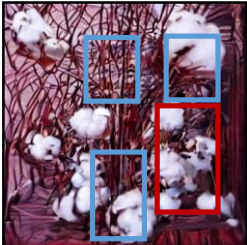
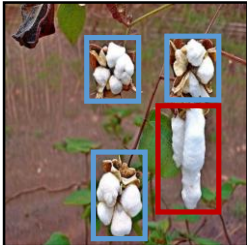
$$\mathcal{L}_{recon} = \mathbb{E}_{x,t,c_t,c_{tl},c_{il},\epsilon \sim \mathcal{N}(0,1)} [\|\epsilon - \epsilon_{\theta}(x_t, t, c_t, c_{tl}, c_{il})\|^2]$$

$$\mathcal{L}_{control} = \mathcal{L}_{recon} + \gamma \mathcal{L}_{recon} \odot \mathcal{M}$$

Dataset Synthesis Pipeline



Qualitative Results

	Annotation	ReCo	GLIGEN	ControlNet	GeoDiffusion	ODGEN
MRI Image						
Apex Game						
Underwater						
Cotton						

Quantitative Results

Table 1: FID (\downarrow) scores computed over 5000 images synthesized by each approach on RF7 datasets. ODGEN achieves better results than the other on all 7 domain-specific datasets.

Datasets	ReCo	GLIGEN	ControlNet	GeoDiffusion	ODGEN
Apex Game	88.69	125.27	97.32	120.61	58.21
Robomaster	70.12	167.44	134.92	76.81	57.37
MRI Image	202.36	270.52	212.45	341.74	93.82
Cotton	108.55	89.85	196.87	203.02	85.17
Road Traffic	80.18	98.83	162.27	68.11	63.52
Aquarium	122.71	98.38	146.26	162.19	83.07
Underwater	73.29	147.33	126.58	125.32	70.20

Table 2: mAP@.50:.95 (\uparrow) of YOLOv5s / YOLOv7 on RF7. Baseline models are trained with 200 real images only, whereas the other models are trained with 200 real + 5000 synthetic images from various methods. ODGEN leads to the biggest improvement on all 7 domain-specific datasets.

	Baseline	ReCo	GLIGEN	ControlNet	GeoDiffusion	ODGEN
real + synth #	200 + 0	200 + 5000	200 + 5000	200 + 5000	200 + 5000	200 + 5000
Apex Game	38.3 / 47.2	25.0 / 31.5	24.8 / 32.5	33.8 / 42.7	29.2 / 35.8	39.9 / 52.6
Robomaster	27.2 / 26.5	18.2 / 27.9	19.1 / 25.0	24.4 / 32.9	18.2 / 22.6	39.6 / 34.7
MRI Image	37.6 / 27.4	42.7 / 38.3	32.3 / 25.9	44.7 / 37.2	42.0 / 38.9	46.1 / 41.5
Cotton	16.7 / 20.5	29.3 / 37.5	28.0 / 39.0	22.6 / 35.1	30.2 / 36.0	42.0 / 43.2
Road Traffic	35.3 / 41.0	22.8 / 29.3	22.2 / 29.5	22.1 / 30.5	17.2 / 29.4	39.2 / 43.8
Aquarium	30.0 / 29.6	23.8 / 34.3	24.1 / 32.2	18.2 / 25.6	21.6 / 30.9	32.2 / 38.5
Underwater	16.7 / 19.4	13.7 / 15.8	14.9 / 18.5	15.5 / 17.8	13.8 / 17.2	19.2 / 22.0

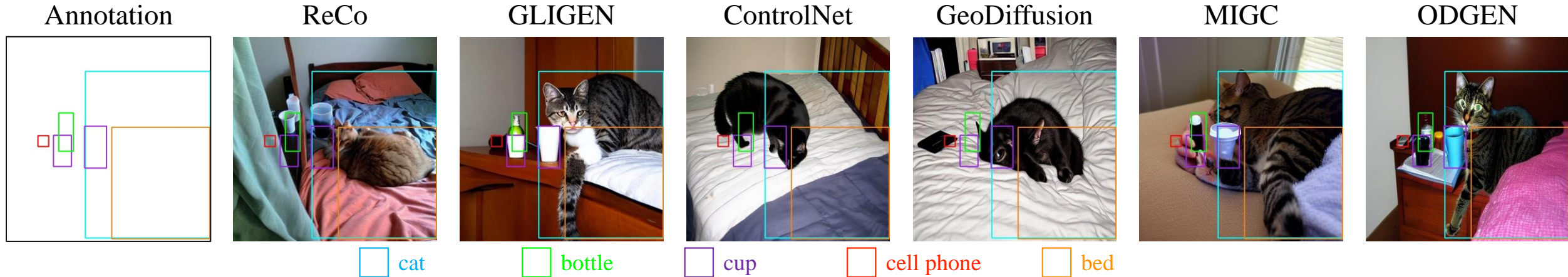
Quantitative Results

Table 3: mAP@.50 / mAP@.50:.95 (\uparrow) results of ODGEN trained on larger-scale datasets of 1000 real images. The top 3 rows show results of YOLOv5s and the bottom 3 rows show results of YOLOv7. Baseline models are trained with 1000 real images only, whereas the other models are trained with 1000 real + 5000 / 10000 synthetic images from various methods. ODGEN leads to more significant improvement than other methods.

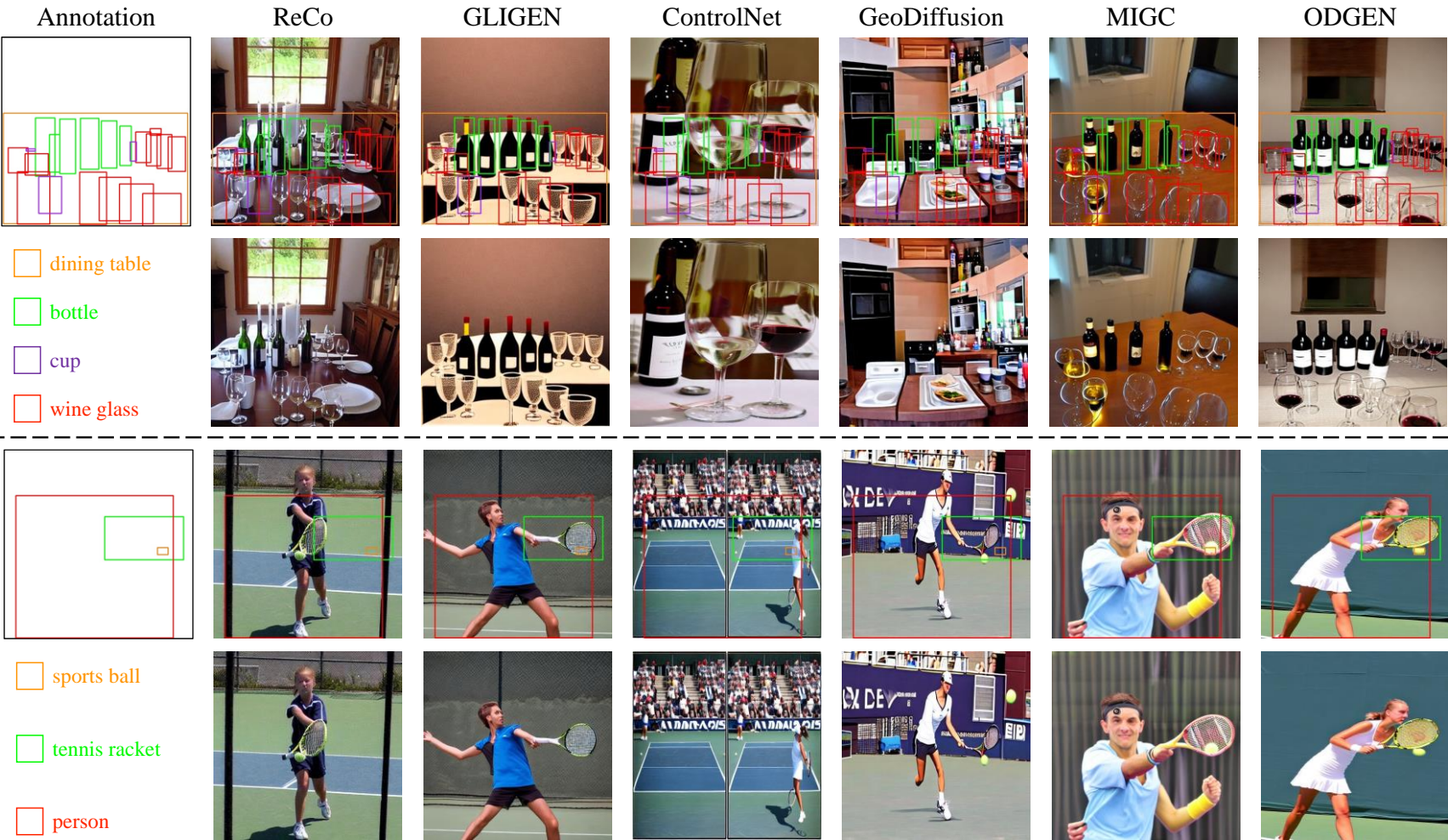
Datasets	Apex Game	Apex Game	Apex Game	Underwater	Underwater	Underwater
real + synth #	1000 + 0	1000 + 5000	1000 + 10000	1000 + 0	1000 + 5000	1000 + 10000
ReCo	83.2 / 53.5	78.7 / 46.9	82.0 / 46.9	55.6 / 29.2	55.1 / 28.4	55.9 / 29.1
GeoDiffusion	83.2 / 53.5	80.0 / 47.2	82.5 / 47.5	55.6 / 29.2	54.2 / 27.9	54.3 / 28.0
ODGEN	83.2 / 53.5	83.3 / 53.5	83.6 / 53.6	55.6 / 29.2	59.6 / 32.5	56.3 / 29.8
ReCo	83.8 / 55.0	80.5 / 50.7	79.2 / 49.9	54.6 / 28.3	56.5 / 28.7	56.4 / 30.1
GeoDiffusion	83.8 / 55.0	81.2 / 51.0	81.0 / 50.5	54.6 / 28.3	57.0 / 28.9	55.8 / 28.9
ODGEN	83.8 / 55.0	84.4 / 55.2	84.0 / 55.0	54.6 / 28.3	58.2 / 29.8	62.1 / 31.8

Experiments on General Domain

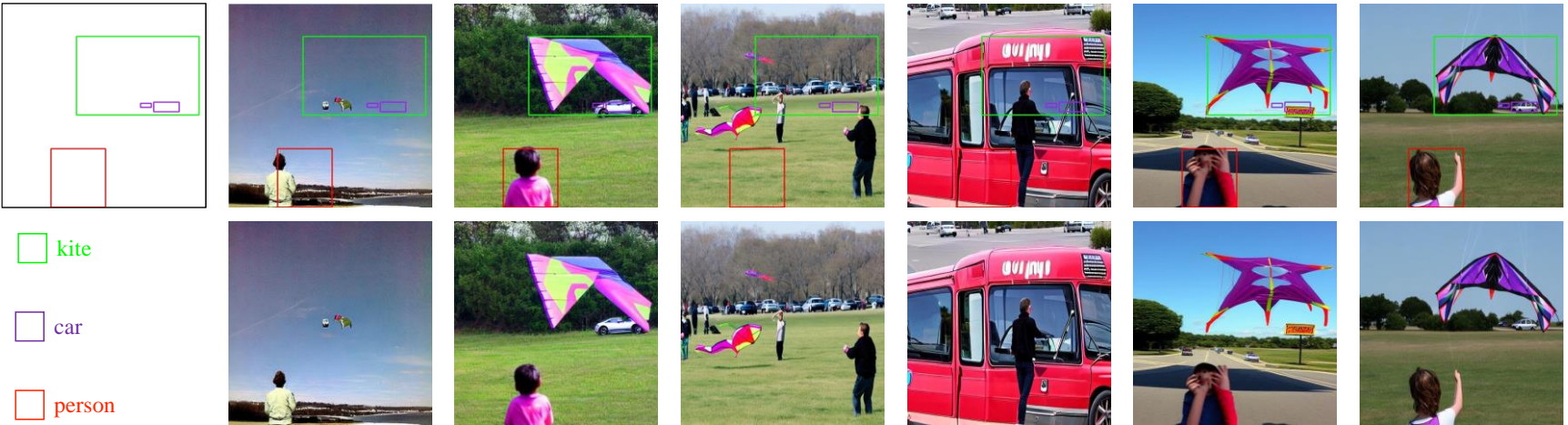
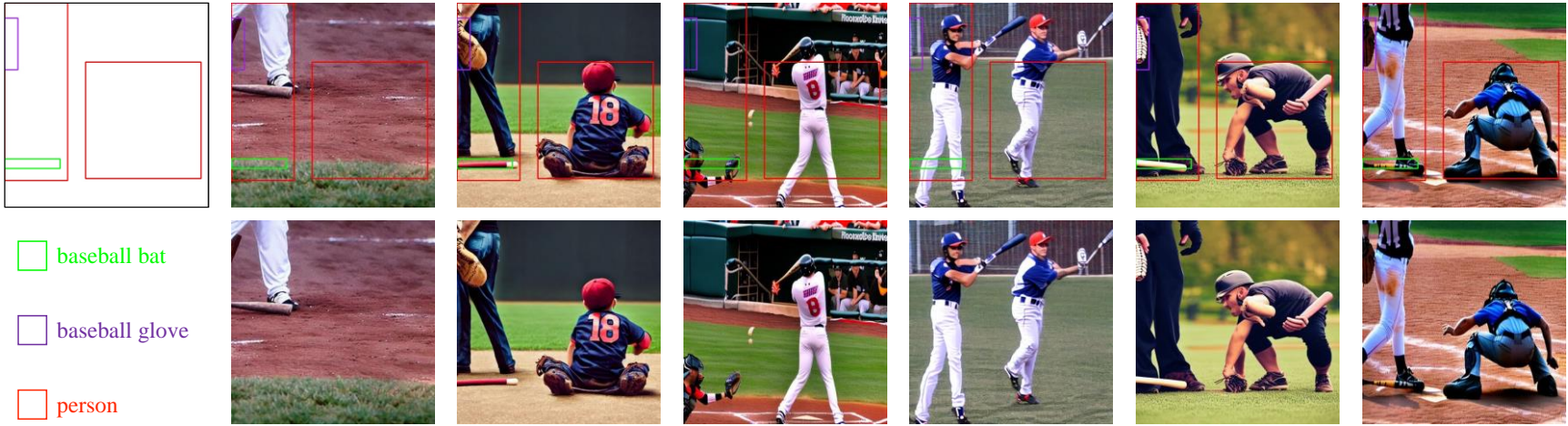
- Training: COCO-2014 train split
- Test: COCO-2014 val split



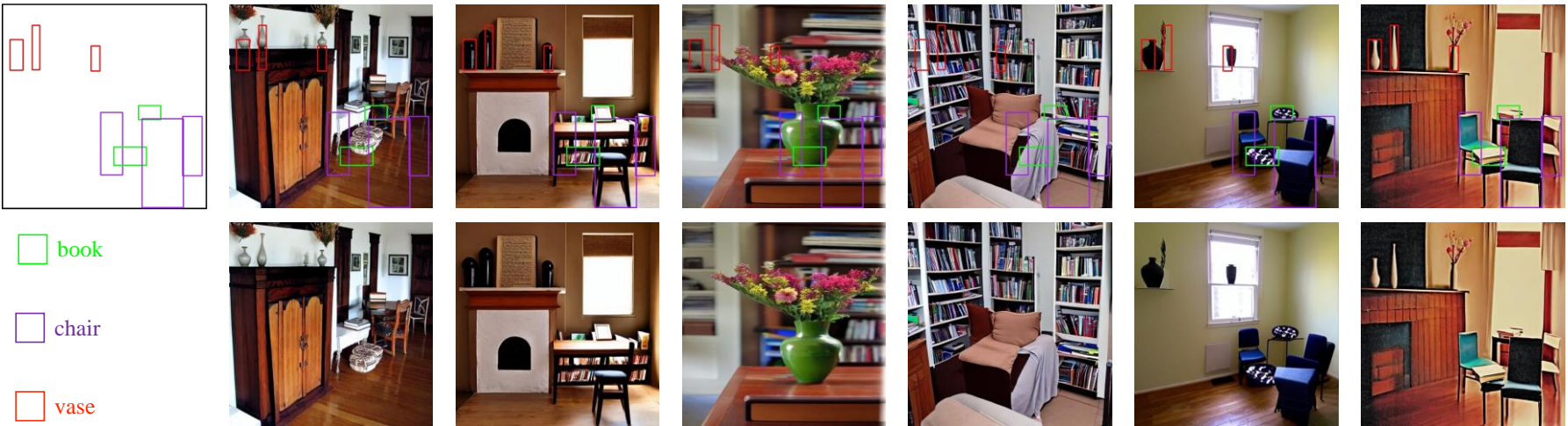
Qualitative Results



Qualitative Results



Qualitative Results



Quantitative Results

Table 4: FID (\downarrow) and mAP (\uparrow) of YOLOv5s / YOLOv7 on COCO. FID is computed with 41k synthetic images. For mAP, YOLO models are trained from scratch on 10k synthetic images and validated on 31k real images. ODGEN outperforms all the other methods in terms of both fidelity and trainability.

Metrics	ReCo	GLIGEN	ControlNet	Geo-Diffusion	MIGC	Instance-Diffusion	ODGEN
FID	18.36	26.15	25.54	30.00	21.82	23.29	16.16
mAP@.50	7.60 / 11.01	6.70 / 9.42	1.43 / 1.15	5.94 / 9.21	9.54 / 16.01	10.00 / 17.10	18.90 / 24.40
mAP@.50:.95	3.82 / 5.29	3.56 / 4.60	0.52 / 0.38	2.37 / 4.44	4.67 / 8.65	5.42 / 10.20	9.70 / 14.20

Conclusions

- Propose to fine-tune pre-trained diffusion models with both cropped foreground patches and entire images to generate high-quality domain-specific target objects and background scenes.
- Design a novel strategy to control diffusion models with object-wise text prompts and synthetic visual conditions, improving their capability of generating and controlling complex scenes.
- Conduct extensive experiments to demonstrate that our synthetic data effectively improves the performance of object detectors and outperforms prior works.