Adversarial Supervision Makes Layout-to-Image Diffusion Models Thrive

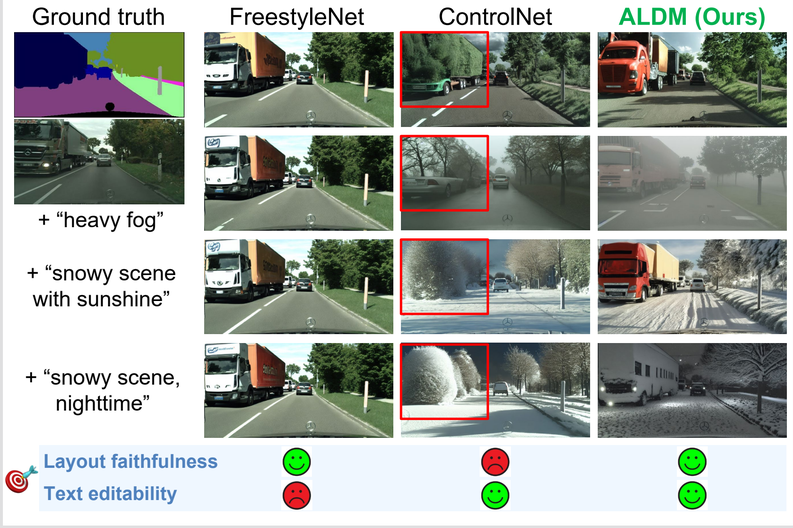
Yumeng Li[1,2], Margret Keuper[2,3], Dan Zhang[1,4], Anna Khoreva[1]

[1]Bosch Center for Artificial Intelligence, [2]University of Mannheim, [3]MPI for Informatics, [4]University of Tübingen

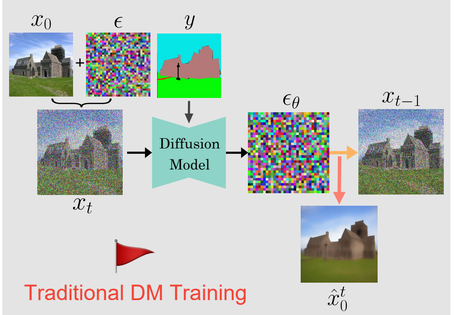
🤗 Code & Models!

1. Motivation

- Adopt **pretrained** large scale Text-to-Image diffusion models, e.g., Stable Diffusion, for Layout-to-Image task.
- We identify two challenges:
  1. **Alignment** with the desired layout condition
  2. **Editability** via text control



2. Traditional Training



Randomly sampled **single** step

- Sample the noisy latent x_t at a random timestep

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon$$

MSE reconstruction loss

- Learn to denoise, i.e., predict the added noise ϵ

$$\mathcal{L}_{noise} = \mathbb{E}_{\epsilon\sim N(0,I),y,t}\left[\left\|\epsilon - \epsilon_\theta(x_t, y, t)^2\right\|\right]$$

3. Novel Training Strategy

Surpassing ControlNet

Model Agnostic Training Strategy

Our ALDM

Adversarial Supervision

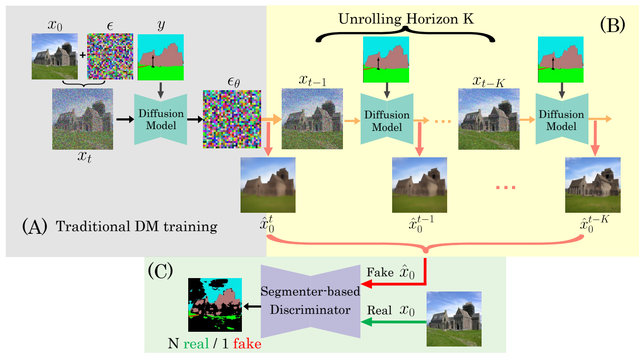
Explicitly leverage the label map condition for supervision

Multistep Unrolling

Better imitate the inference time sampling with more comprehensive supervision

Encourage the **consistent adherence** to the given label map **over a time horizon**

4. Method Overview



(B) Multistep unrolling:

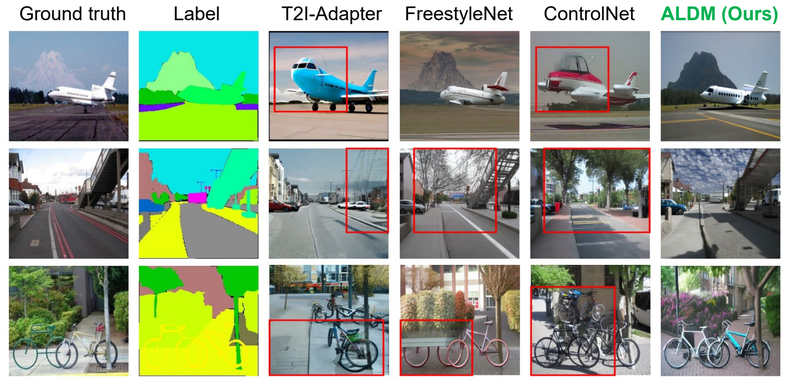
- Bridge the gap between inference time sampling and single timestep sampling during training
- Employ supervision over consecutive denoising steps
  → **Consistent alignment** with the given layout condition

(C) Adversarial supervision:

- (N+1)-Classes-Segmenter-Based **Discriminator**
  Real → N semantic classes, Fake → One extra "fake" class
- **Generator** (diffusion model) should learn to fool the discriminator, i.e., synthesize samples that well comply with the label map

5. Results

ALDM can effectively enhance the layout faithfulness!



Metric:

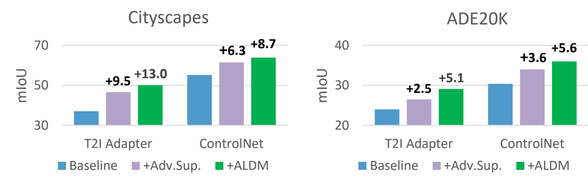
- **mIoU**: measure alignment with the layout condition
- **TIFA**: measure text editability

➢ By default, ALDM represents ControlNet + Adv. Supervision + Multistep unrolling.

Method	Cityscapes		ADE20K	
	mIoU ↑	TIFA ↑	mIoU ↑	TIFA ↑
FreestyleNet	68.8	0.300	36.1	0.740
T2I-Adapter	37.1	0.902	24.0	0.892
ControlNet	55.2	0.822	30.4	0.838
ALDM (Ours)	63.9	0.856	36.0	0.888

6. Ablation Study

The proposed Adversarial Supervision & Multistep Unrolling can effectively boost different layout-to-image diffusion models, e.g., T2I-Adapter and ControlNet.



7. Application: Improved Domain Generalization