amazon web services

澳門大學 UNIVERSIDADE DE MACAU UNIVERSITY OF MACAU

科技學院 Faculdade de Ciências e Tecnologia Faculty of Science and Technology

Rethinking The Training And Evaluation of Rich-Context Layout-to-Image Generation

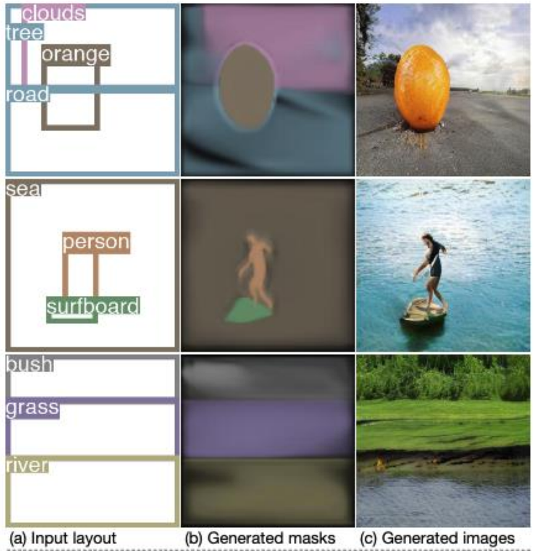
Jiaxin Cheng, Zixu Zhao, Tong He, Tianjun Xiao, Yicong Zhou, Zheng Zhang

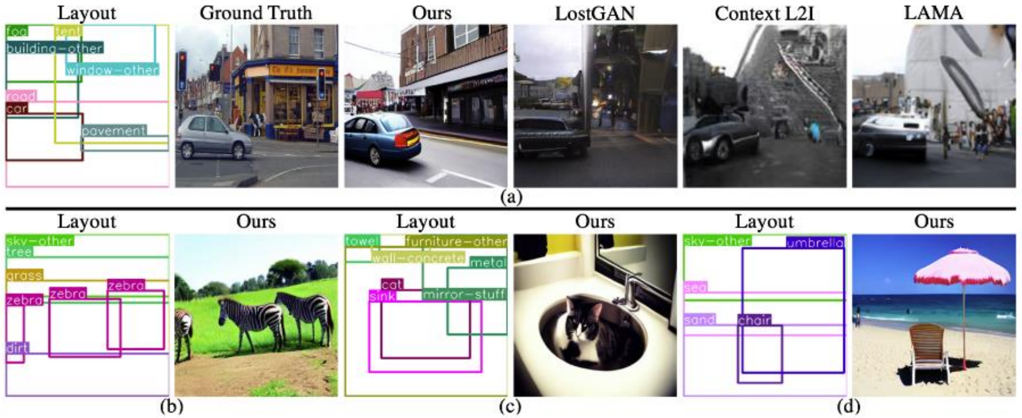
Amazon Web Services Shanghai AI Lab

University of Macau

Layout-to-Image Generation

- Given layout bounding boxes and instance description, generating an image complies the layout and description





GAN-based Models, closed-set

Diffusion-based Models, closed-set

Diffusion-based Models, open-set

Improve generation quality

Involve larger set of descriptions

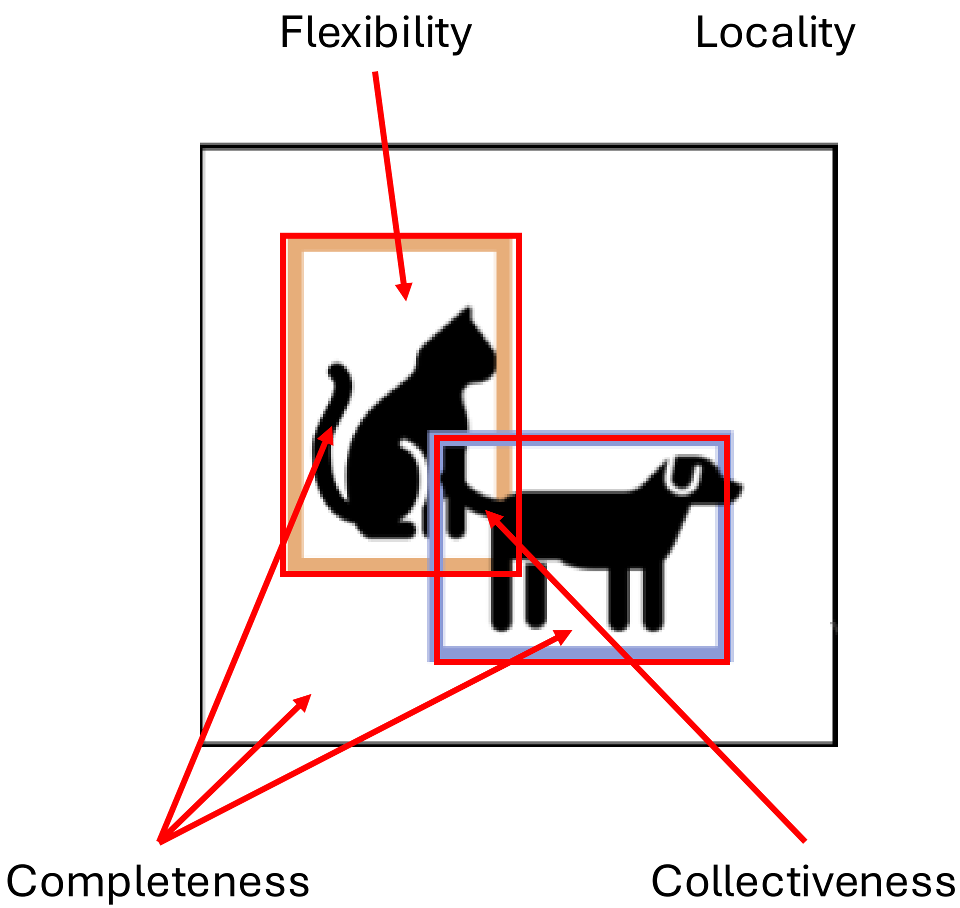
Image Synthesis from Layout with Locality-Aware Mask Adaption
LayoutDiffuse: Adapting Foundational Diffusion Models for Layout-to-Image Generation
GLIGEN: Open-Set Grounded Text-to-Image Generation

Rich-context Layout2image Generation

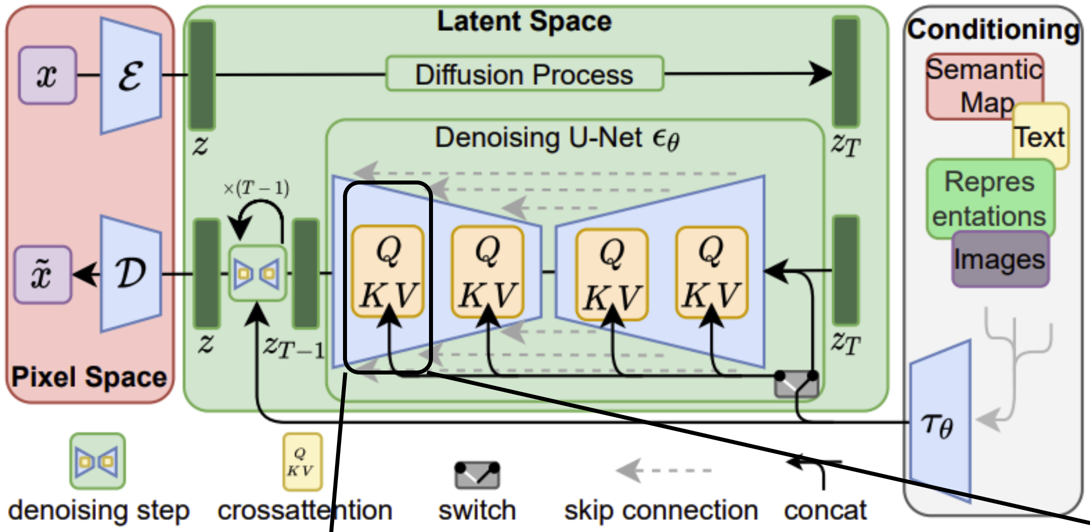
Rich-context: The description for each object is more complex and lengthier.

Desired Properties of L2I

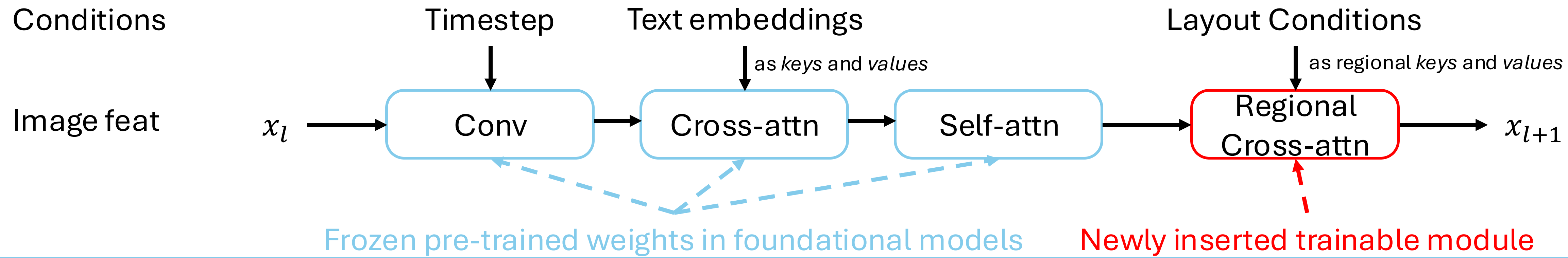
- Flexibility: The model must accurately understand rich-context descriptions
- Locality: Generated object should be bounded within its layout bbox
- Completeness: All region should be treated equally when adding layout conditions, including background
- Collectiveness: all object should be considered for overlapping region



Where/How to Insert Layout Information?

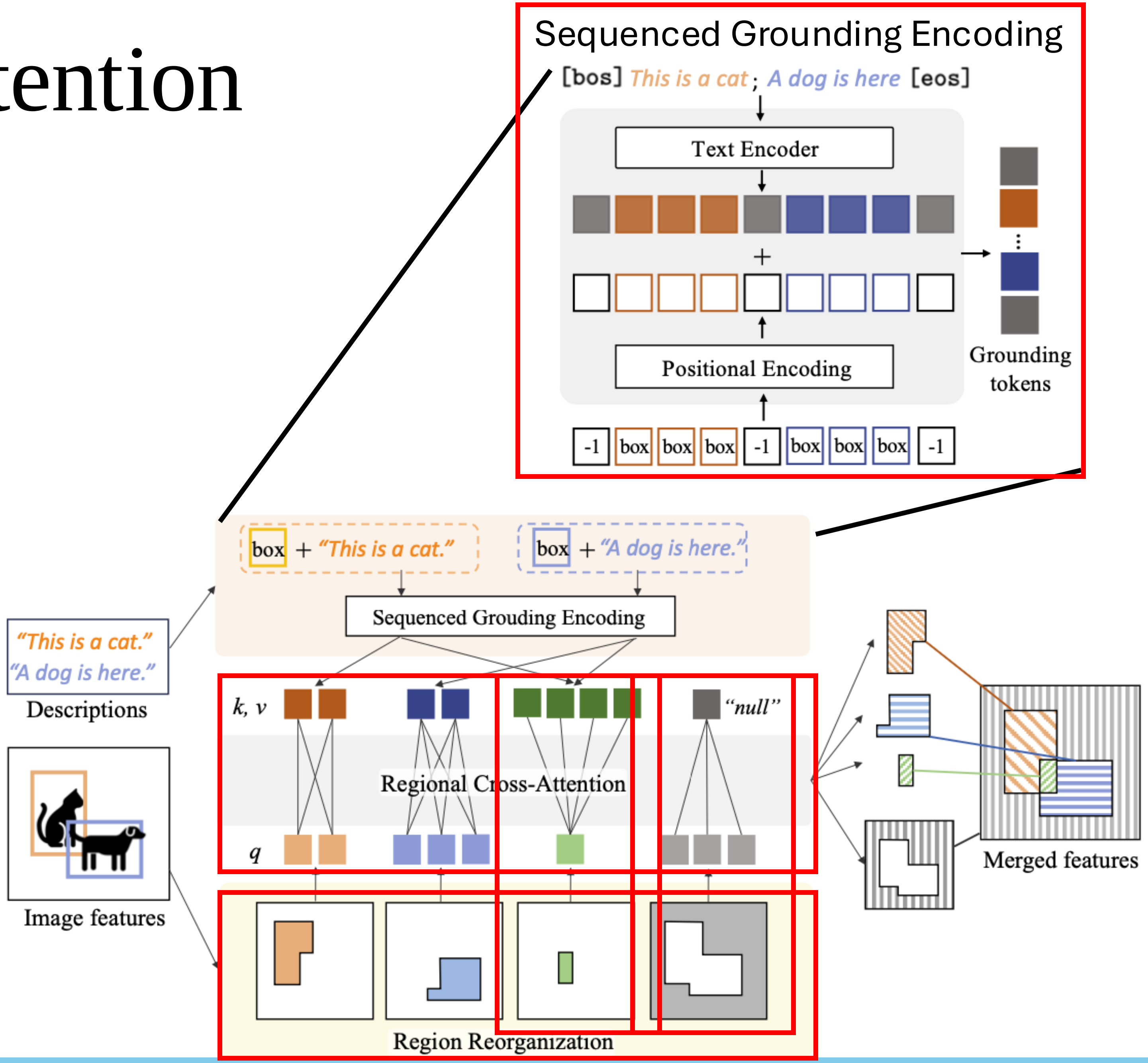


- Each model block is composed of conv, cross-attn and self-attn layers.
- The textual information is inserted in cross-attention layers as *keys* and *values*
- We insert a regional cross-attn layer with layout conditions after each self-attn layer.
- Each object description cross-attends with regional image feature as *keys* and *values*



Regional Cross-Attention

- We partition the object regions according to their overlapping states, naming region reorganization. (Locality)
- We apply cross-attention between visual and textual tokens within each re-partitioned region. (Flexibility)
- Overlapping region will cross-attend with grounding tokens of all objects within it. (Collectiveness)
- The background will attend with a learnable null-token. (Completeness)
- The grounding tokens are composed of textual tokens and location tokens. (For model to recognize overlapped objects with identical descriptions)



Training Setting

Loss function

$$x_t(x_0, t)$$

Noisy image $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$

$$L = E_{t,\varepsilon,x_0}\left[\left(\varepsilon - \varepsilon_\theta(x_t(x_0,t),t,l)\right)^2\right]$$

Predicting the noise ε added on the image

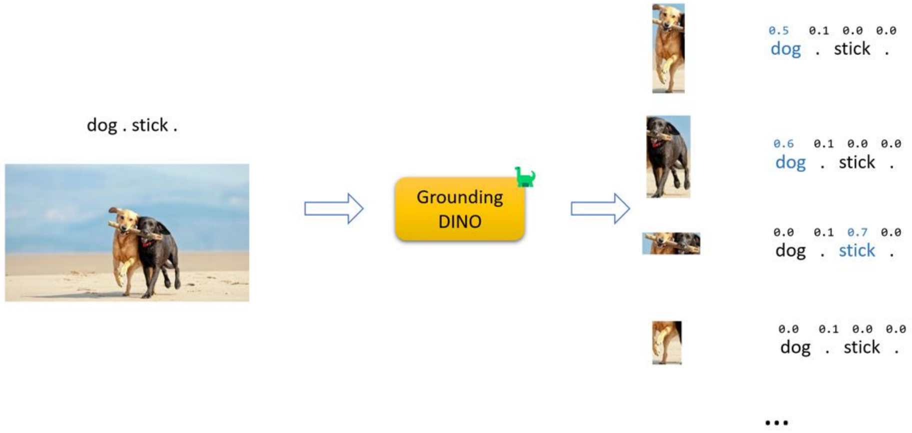
$$\varepsilon_\theta(x_t, t, l)$$

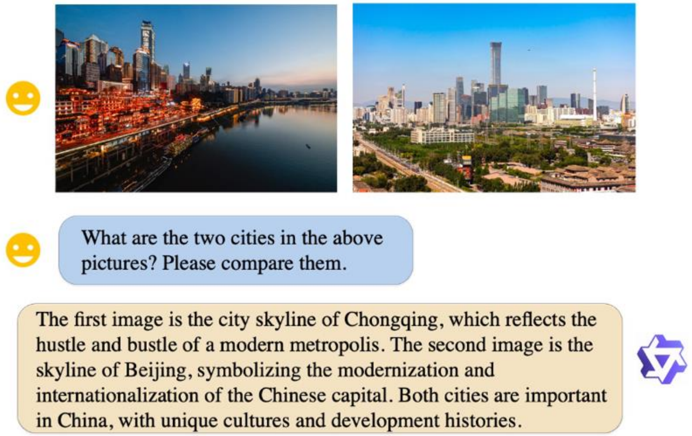
Denoising model conditioned on the noisy image x_t, timestep t and layout information l

---

Dataset Generation

Recognizing	Locating	Labelling
Recognize Anything, image tagging	GroundingDINO, open-set object detection	Qwen-VL, object description



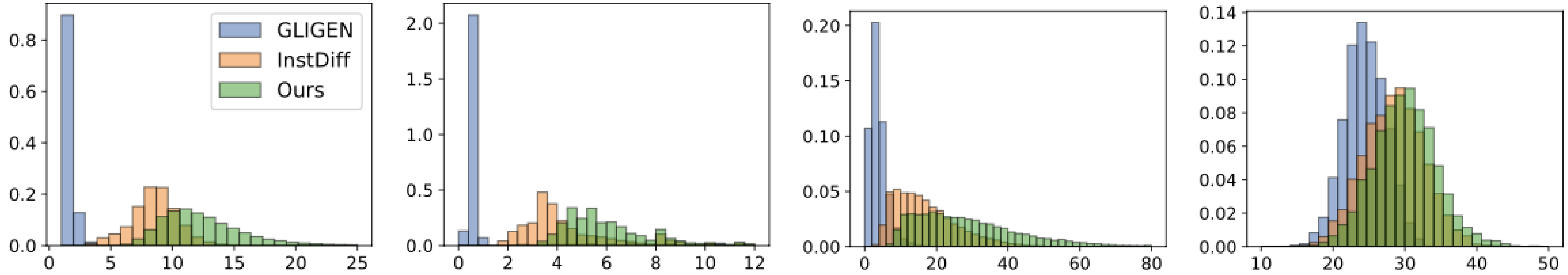


Recognize Anything: A Strong Image Tagging Model
Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection
Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond

Rich-context Dataset

Image → Recognize Anything Model → obj1 category, obj2 category, obj3 category, obj4 category, … → Grounding DINO → obj1 bbox, obj2 bbox, obj3 bbox, obj4 bbox, … → Qwen Visual-LLM → obj1 desc, obj2 desc, obj3 desc, obj4 desc, …

Synthetic rich-context dataset generation pipeline



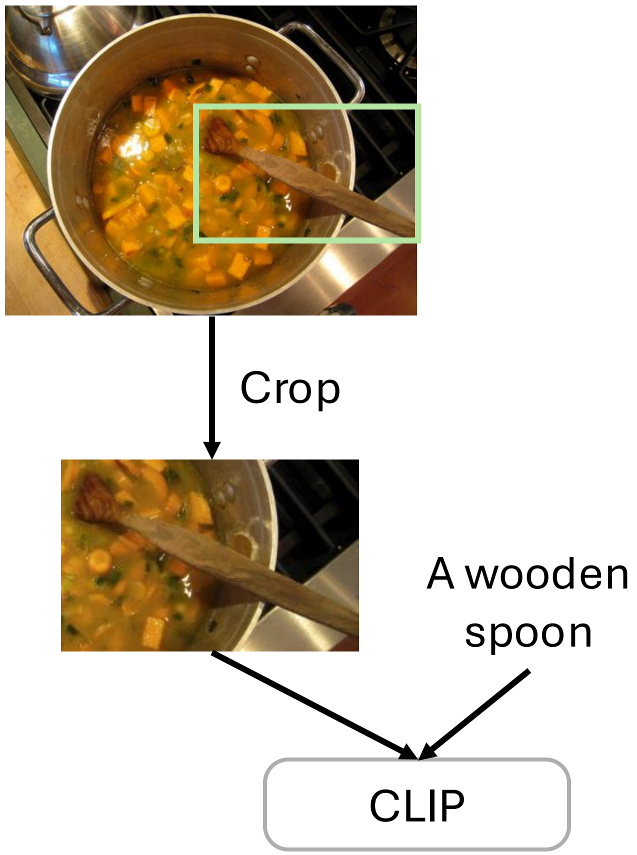
(a) Average Caption Length (b) Gunning Fog Score (Complexity) (c) Unique Words/Sample (Diversity) (d) Object-label CLIP Alignment Score

Statistics of synthetic rich-context dataset

GLIGEN: Open-Set Grounded Text-to-Image Generation
InstanceDiffusion: Instance-level Control for Image Generation

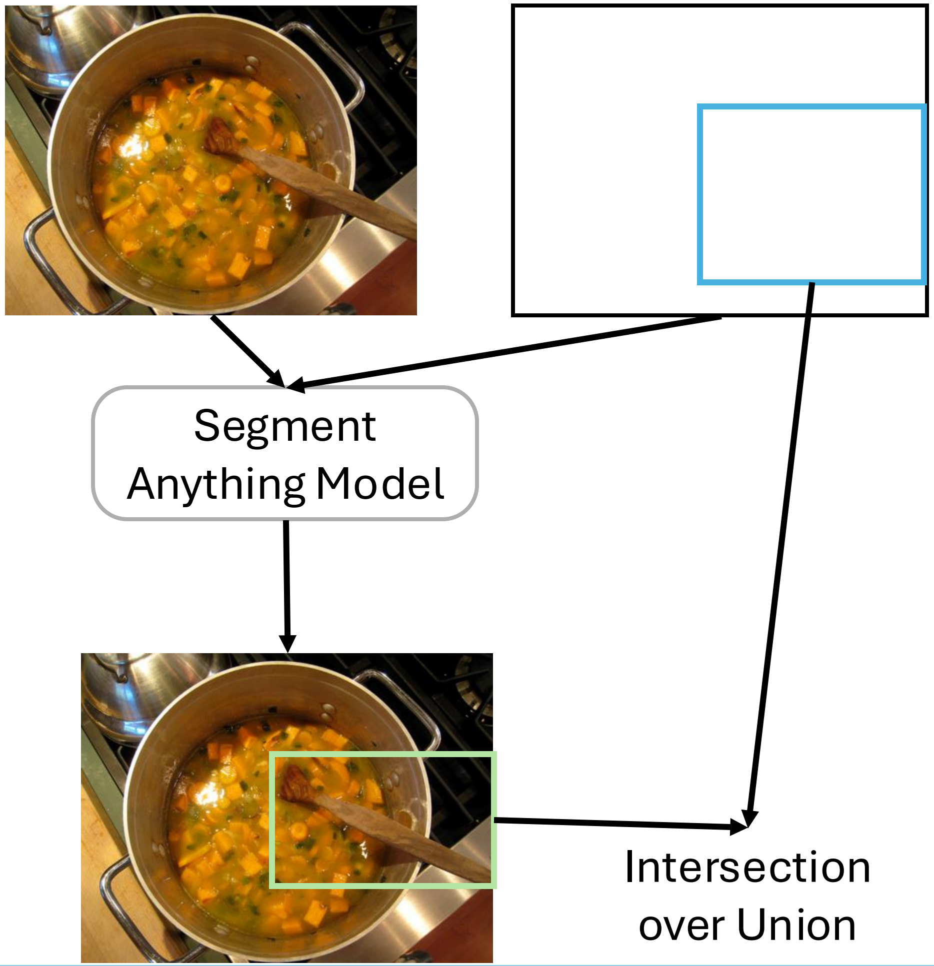
Evaluation Metric for Rich-context L2I

Crop CLIP score

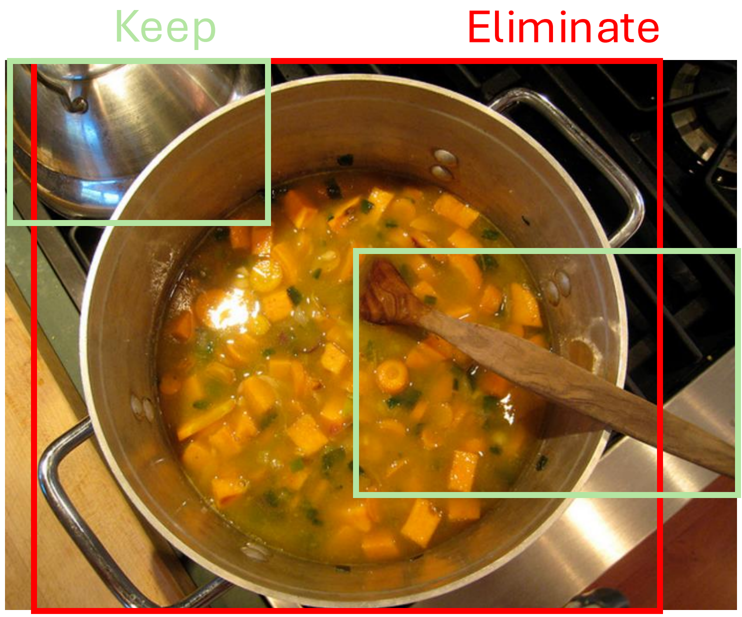


CLIP similarity

SAM IoU score



Intersection over Union

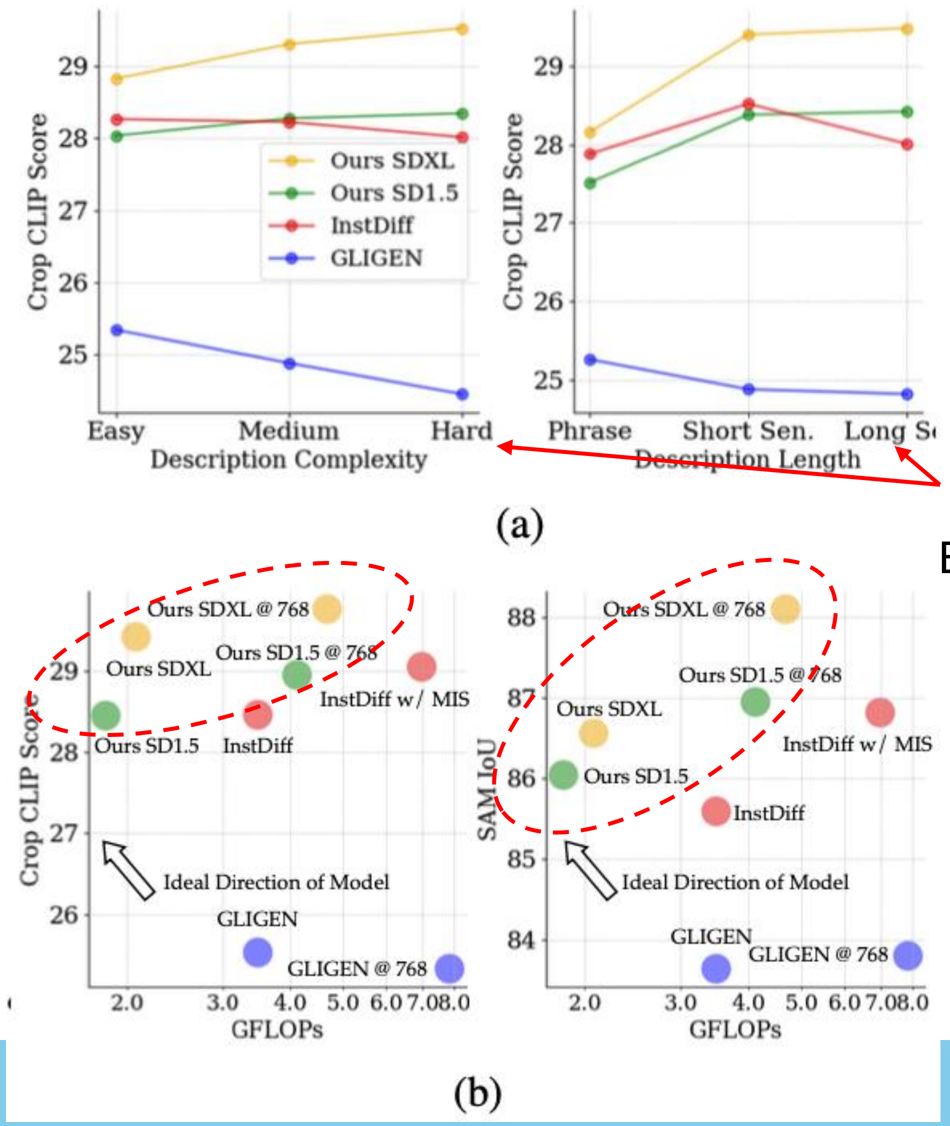


Eliminating results (during evaluation) that do not align well with human perspective

- Conduct a user study for object-text alignment and layout fidelity
- Object size < 5% and >50% of image size not align well with human feedback

Segment Anything
Learning Transferable Visual Models From Natural Language Supervision

Performance Comparison



Better for complex and lengthy descriptions

Better performance-computation trade off

- Figure (a), our method shows better performance when the complexity or length of object caption increases
- Figure (b), our method has a better performance-computation trade-off

Ablation Studies

Backbone	Dataset		Attention Module			CropCLIP	SAMIoU
	Word/Phrase	Rich-context	SelfAttn GLIGEN	SelfAttn InstDiff	CrossAttn Ours		
SDXL	✓				✓	25.40	86.76
SDXL		✓			✓	29.79	88.10
SD1.5		✓	✓			25.56	82.72
SD1.5		✓		✓		28.36	85.58
SD1.5		✓			✓	28.94	86.91

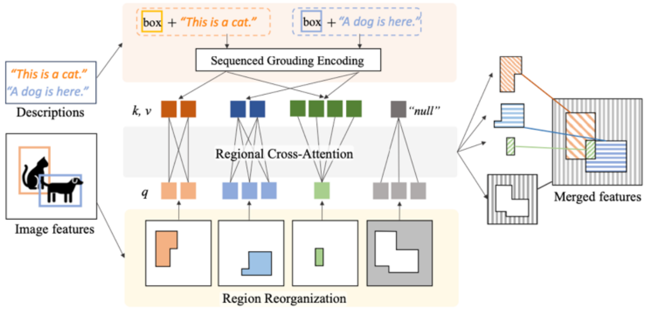
- Word-level dataset trained L2I model can hardly generalize to the rich-context descriptions.
- The regional cross-attention module is more suitable for rich-context L2I than existing self-attention-based layout conditioning module.

Summary

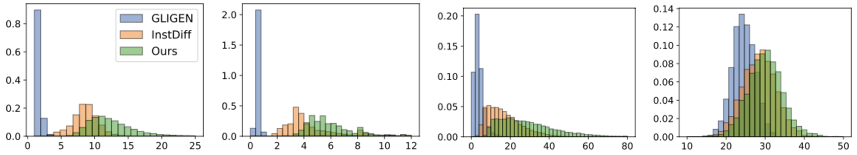
Scan me!

Project page

Model

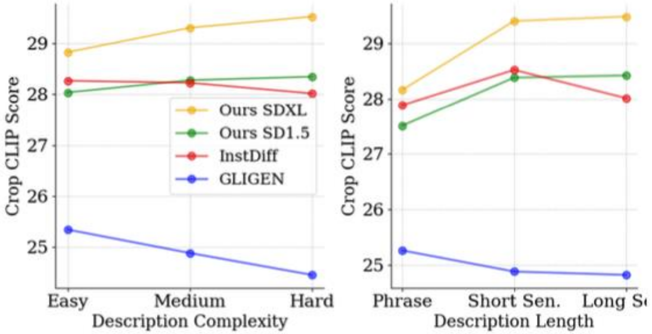


Dataset



(a) Average Caption Length (b) Gunning Fog Score (Complexity) (c) Unique Words/Sample (Diversity) (d) Object-label CLIP Alignment Score

Evaluation



- **A fine-tuned layout-to-image model** established on foundational diffusion model
- **Propose regional cross-attention** to improve the layout-to-image generation quality on rich-context descriptions

- **A synthetic dataset** curated with three large pre-trained multi-modality models
- **Rich-context annotations**: the annotations are more diverse, complex and lengthy while align better with object

- **Propose two metrics** for rich-context object-text alignment and layout fidelity
- The proposed method **performs better on complex and lengthy descriptions**