



# STRUCT Group Paper Reading

## **InstanceDiffusion: Instance-level Control for Image Generation**

Xudong Wang<sup>1,2</sup> Trevor Darrell<sup>2</sup> Sai Saketh Rambhatla<sup>1</sup> Rohit Girdhar<sup>1</sup> Ishan Misra<sup>1</sup>  
<sup>1</sup>GenAI, Meta      <sup>2</sup>UC Berkeley

code: <https://github.com/frank-xwang/InstanceDiffusion/>

CVPR 2024

PRESENTER: XIANG GAO

2024/10/20

---

## **2 Paper Reading Content Outline**

---

# **Content**

---

**Background/ 04**

**Author/ 19**

**Method/ 25**

**Experiment/ 38**

---

---

### **3 Paper Reading Content Outline**

---

# Content

---

**Background/ 04**

**Author/ 19**

**Method/ 25**

**Experiment/ 38**

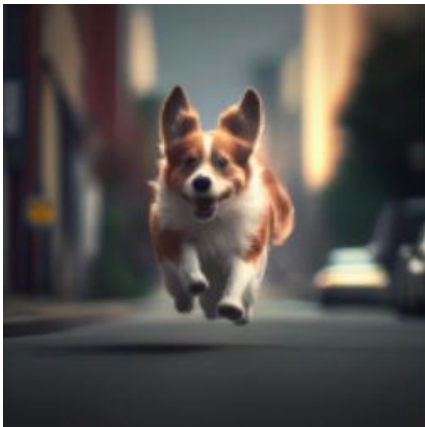
---

---

## 4 Background—global-level T2I control

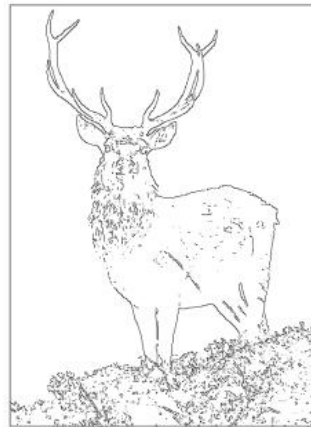
---

T2I diffusion model



“a running dog”

Spatial control by ControlNet

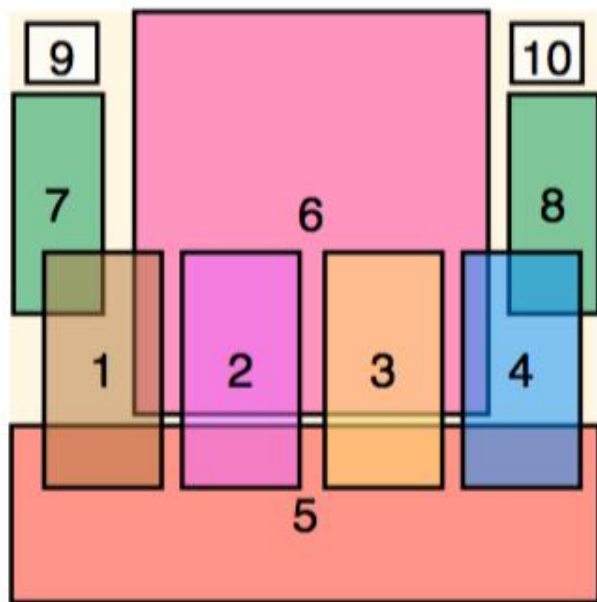


“giant deer”

Lacking the ability of instance-level control !

## 5 Background—instance-level T2I control

### Preview of instance-level control effect of InstanceDiffusion



**Image Caption:** An image depicting in the morning. A **brown** cute teddy bear, a **purple** cute teddy bear, a **yellow** cute teddy bear, a **blue** cute teddy bear all standing side by side on a **red** brick road. The scene should be set in front of **Pink** Castle with clear **blue** sky overhead, punctuated by fluffy **white** clouds, and trees with **green** leaves. The **pink** castle should loom majestically in the background. **Instance Captions:** 1-4) A **brown/purple/yellow/blue** teddy bear; 5) a **red** brick road; 6) **Pink** Castle; 7-8) **green** leaves; 9-10) fluffy **white** clouds



# 6 Background—SpaText

## SpaText: Spatio-Textual Representation for Controllable Image Generation

Omri Avrahami<sup>1,2</sup> Thomas Hayes<sup>1</sup> Oran Gafni<sup>1</sup> Sonal Gupta<sup>1</sup>  
Yaniv Taigman<sup>1</sup> Devi Parikh<sup>1</sup> Dani Lischinski<sup>2</sup> Ohad Fried<sup>3</sup> Xi Yin<sup>1</sup>  
<sup>1</sup>Meta AI    <sup>2</sup>The Hebrew University of Jerusalem    <sup>3</sup>Reichman University

CVPR 2023

“in the forest”



“a black cat with a red sweater and a blue jeans”

“on the moon”



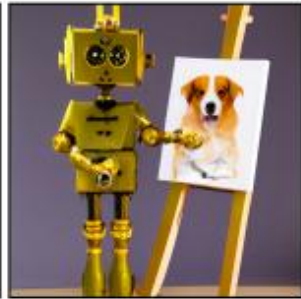
“an astronaut”  
“a horse”

“in the style of  
The Starry Night”



“a black horse”  
“a red full moon”

“in an empty room”



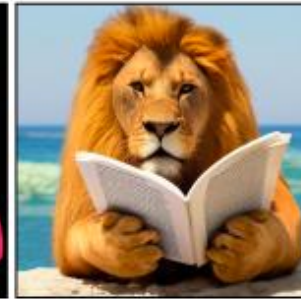
“a canvas with a painting  
of a Corgi dog”  
“a metallic yellow robot”

“on a snowy day”



“a mouse”  
“boxing gloves”  
“a black punching bag”

“at the beach”



“a lion”  
“a book”

## 7 Background—SpaText

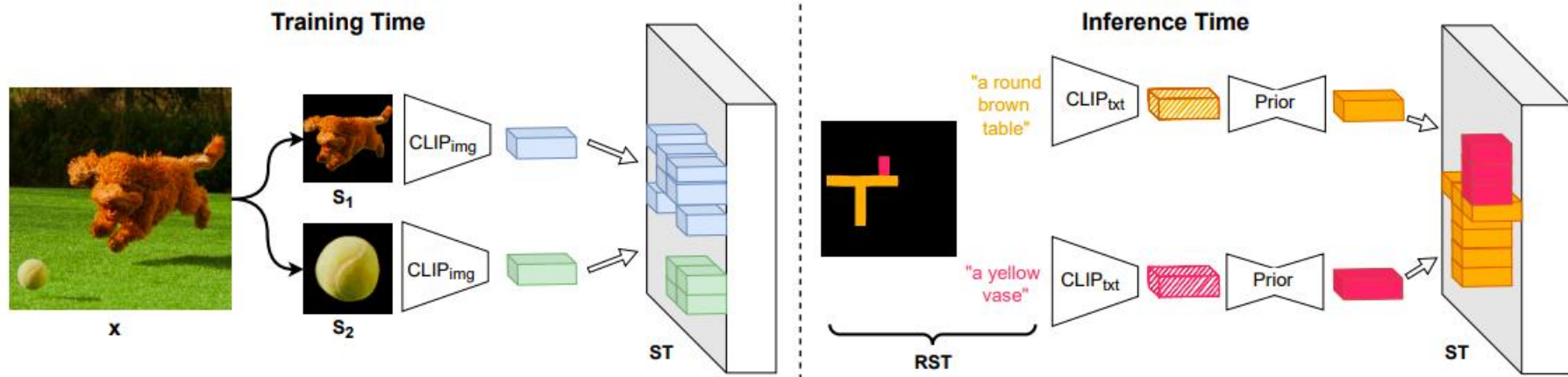


Figure 3. **Spatio-textual representation:** During training (left) — given a training image  $x$ , we extract  $K$  random segments, pre-process them and extract their CLIP image embeddings. Then we stack these embeddings in the same shapes of the segments to form the spatio-textual representation  $ST$ . During inference (right) — we embed the local prompts into the CLIP text embedding space, then convert them using the prior model  $P$  to the CLIP image embeddings space, lastly, we stack them in the same shapes of the inputs masks to form the spatio-textual representation  $ST$ .

$$L_{\text{simple}} = E_{t, x_0, \epsilon} [||\epsilon - \epsilon_{\theta}(x_t, \text{CLIP}_{\text{img}}(x_0), ST, t)||^2]$$

### **GLIGEN: Open-Set Grounded Text-to-Image Generation**

Yuheng Li<sup>1§</sup>, Haotian Liu<sup>1§</sup>, Qingyang Wu<sup>2</sup>, Fangzhou Mu<sup>1</sup>, Jianwei Yang<sup>3</sup>, Jianfeng Gao<sup>3</sup>,  
Chunyuan Li<sup>3¶</sup>, Yong Jae Lee<sup>1¶</sup>

<sup>1</sup>University of Wisconsin-Madison <sup>2</sup>Columbia University <sup>3</sup>Microsoft

<https://gligen.github.io/>

CVPR 2023



## 9 Background—GLIGEN

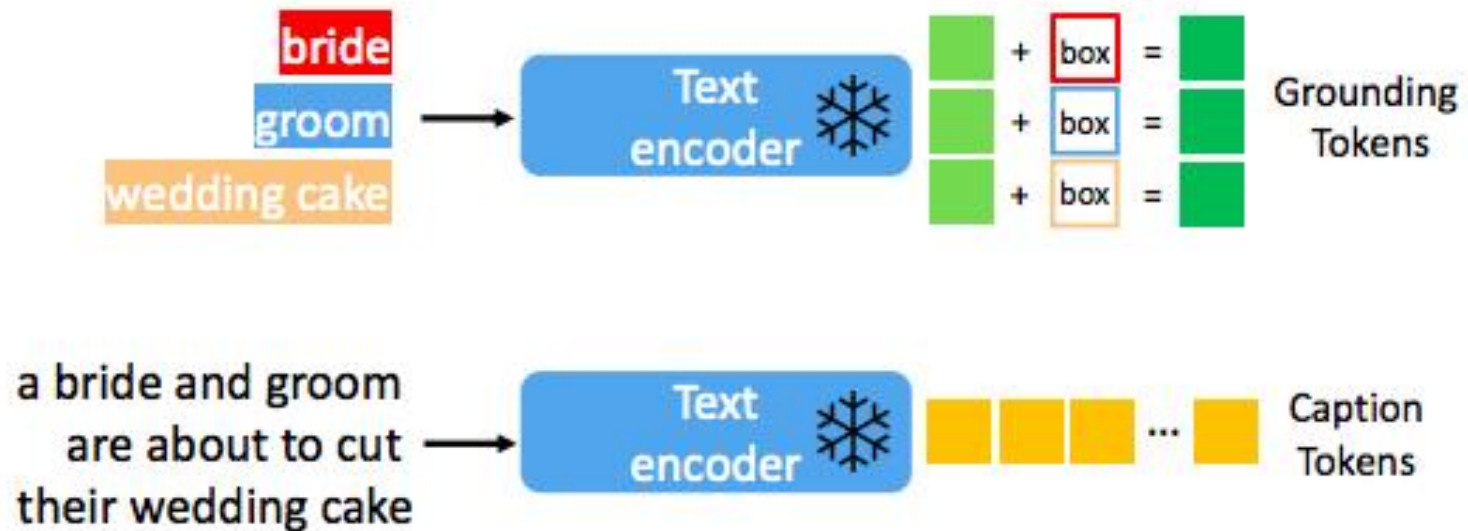


(a) Caption: “A woman sitting in a restaurant with a pizza in front of her ”  
Grounded text: table, pizza, person, wall, car, paper, chair, window, bottle, cup



(b) Caption: “A dog / bird / helmet / backpack is on the grass”  
Grounded image: red inset

# 10 Background—GLIGEN



Instruction:  $\mathbf{y} = (\mathbf{c}, \mathbf{e})$ , with

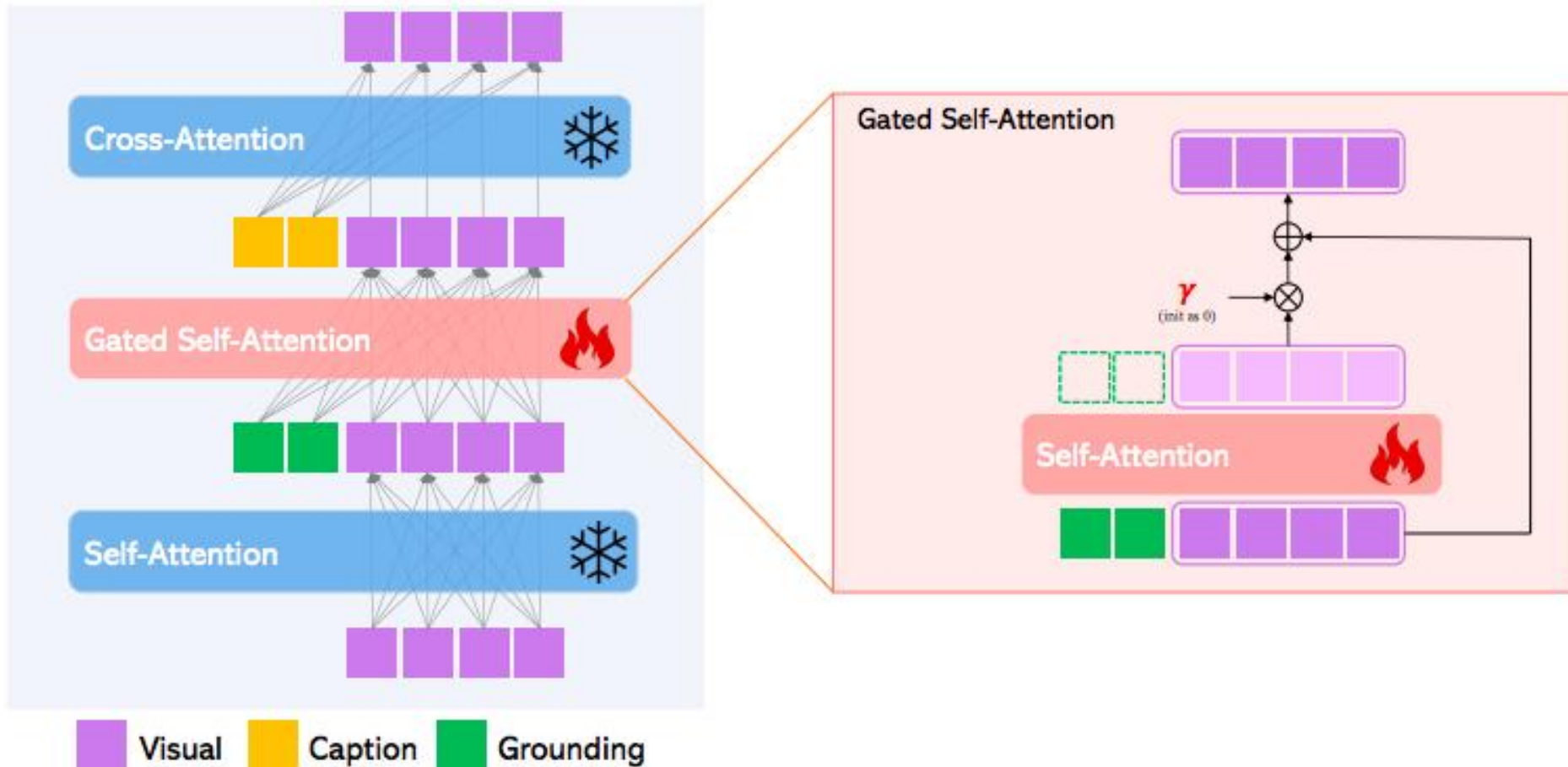
$$h^e = \text{MLP}(f_{\text{text}}(e), \text{Fourier}(\mathbf{l}))$$

Caption:  $\mathbf{c} = [c_1, \dots, c_L]$

Grounding:  $\mathbf{e} = [(e_1, \mathbf{l}_1), \dots, (e_N, \mathbf{l}_N)]$

$$\mathbf{h}^e = [h_1^e, \dots, h_N^e]$$

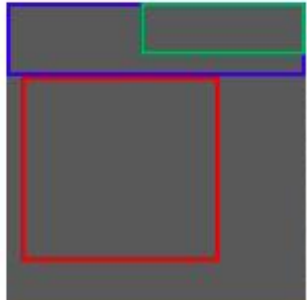
# 11 Background—GLIGEN



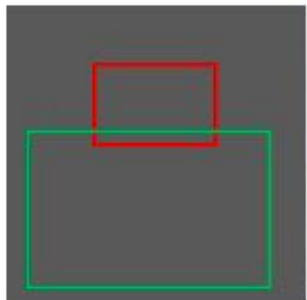
$$\min_{\theta'} \mathcal{L}_{\text{Grounding}} = \mathbb{E}_{\mathbf{z}, \epsilon \sim \mathcal{N}(0, \mathbf{I}), t} [\|\epsilon - f_{\{\theta, \theta'\}}(\mathbf{z}_t, t, \mathbf{y})\|_2^2].$$



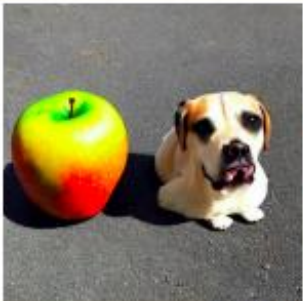
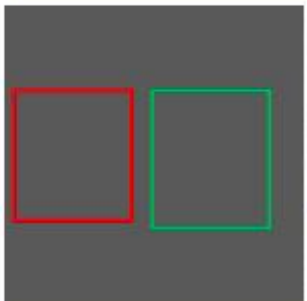
# 12 Background—GLIGEN



Caption: "golden hour, a pekingese is on the beach with an umbrella"  
Grounded text: Pekingese, umbrella, sea



Caption: "a hen is hatching a huge egg"  
Grounded text: hen, egg



Caption: "an apple and a same size dog"  
Grounded text: apple, dog

# 13 Background—BoxDiff

## BoxDiff: Text-to-Image Synthesis with Training-Free Box-Constrained Diffusion

Jinheng Xie<sup>1</sup> Yuexiang Li<sup>2\*</sup> Yawen Huang<sup>2</sup> Haozhe Liu<sup>2,3</sup> Wentian Zhang<sup>2</sup>  
Yefeng Zheng<sup>2</sup> Mike Zheng Shou<sup>1\*</sup>

<sup>1</sup> Show Lab, National University of Singapore <sup>2</sup> Jarvis Lab, Tencent

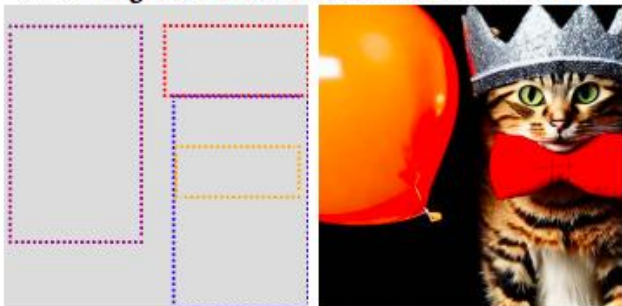
<sup>3</sup> AI Initiative, King Abdullah University of Science and Technology

{sierkinhane, mike.zheng.shou}@gmail.com

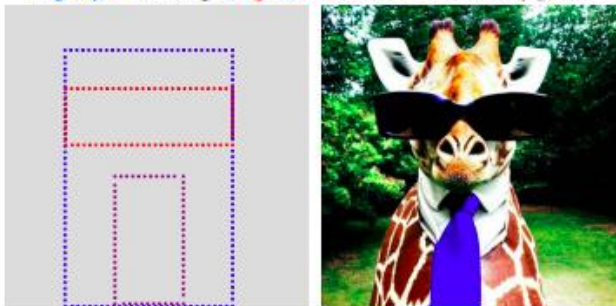
<https://github.com/showlab/BoxDiff>

ICCV 2023

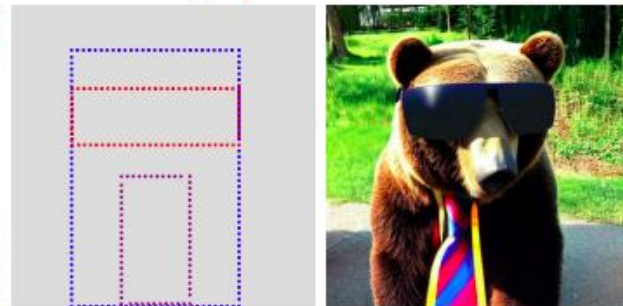
“A smiling *cat*, a *crown*, a *balloon*, and a red *bow*”



“A *giraffe* wearing *sunglasses* and a *tie* looks very proud”



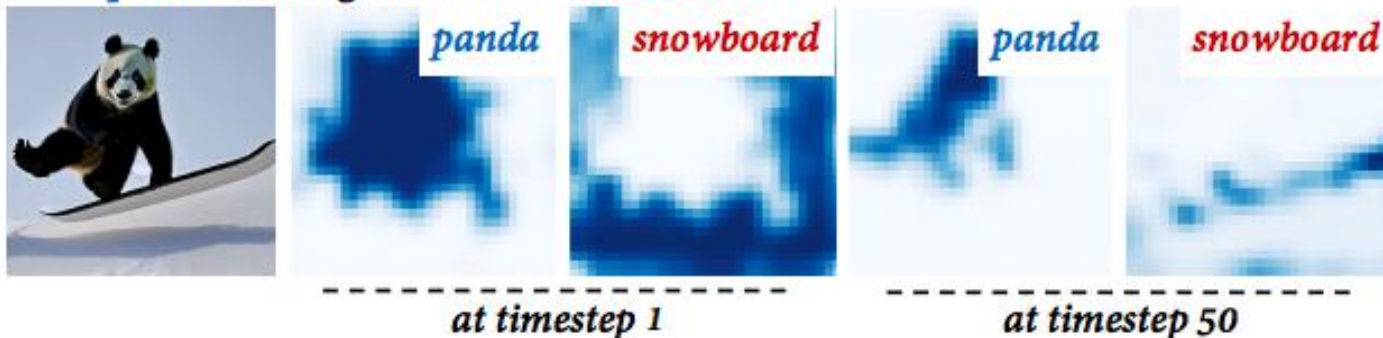
“A *bear* wearing *sunglasses* and a *tie* looks very proud”





# 14 Background—BoxDiff

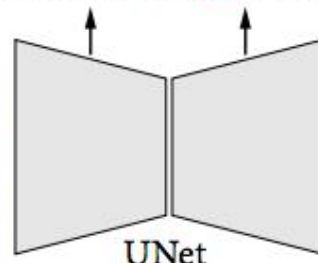
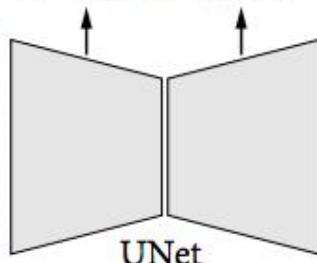
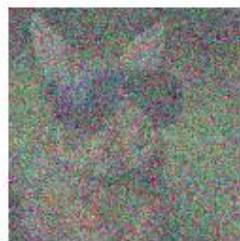
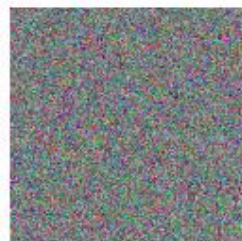
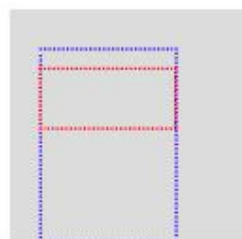
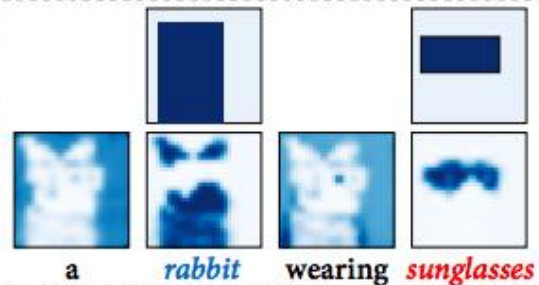
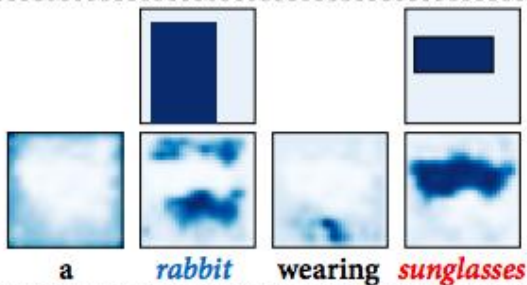
“A panda is doing tricks on the snowboard”



$$\mathcal{L} = \mathcal{L}_{IB} + \mathcal{L}_{OB} + \mathcal{L}_{CC}$$

$$\mathcal{L} = \mathcal{L}_{IB} + \mathcal{L}_{OB} + \mathcal{L}_{CC}$$

apply constraints



$$\mathbf{z}'_t = \mathbf{z}_t - \alpha_t \cdot \nabla \mathcal{L}$$

update latent

$$\mathbf{z}'_{t-1} = \mathbf{z}'_{t-1} - \alpha_{t-1} \cdot \nabla \mathcal{L}$$

update latent

$\mathcal{D}(\cdot)$

---

## 15 Background—DenseDiffusion

---

### **Dense Text-to-Image Generation with Attention Modulation**

Yunji Kim<sup>1</sup>

Jiyoung Lee<sup>1</sup>

Jin-Hwa Kim<sup>1</sup>

Jung-Woo Ha<sup>1</sup>

Jun-Yan Zhu<sup>2</sup>

<sup>1</sup>NAVER AI Lab

<sup>2</sup>Carnegie Mellon University

ICCV 2023

# 16 Background—DenseDiffusion

DenseDiffusion (Ours)

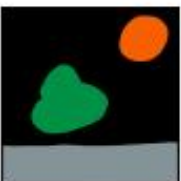
Stable Diffusion [46]



A painting of a couple holding a **yellow umbrella** in a **street on a rainy night**. The woman is wearing a **white dress** and the man is wearing a **blue suit**.



There is a **cute monkey** on a **thick branch** who is holding a **pink rose**. It is on the top of a **huge tree**, and the **sky is so wide and blue**.

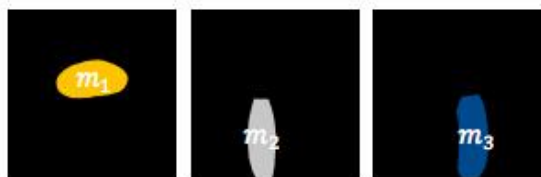


A painting of a **dog riding a flying bicycle**, over a **big city** with a **yellowish full moon** in the **night sky**.

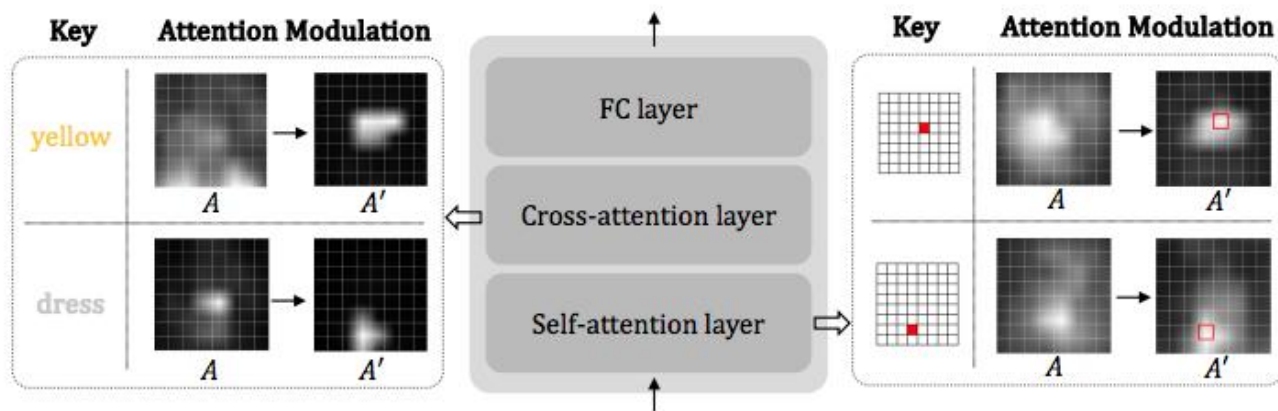


# 17 Background—DenseDiffusion

Dense Caption



A painting of a couple holding a **yellow umbrella** ( $c_1$ ) in a street on a rainy night. The woman is wearing a **white dress** ( $c_2$ ) and the man is wearing a **blue suit** ( $c_3$ ).



Dense Diffusion (Ours)



SD-Pww [4]



Composable Diffusion [36]



Structure Diffusion [17]



Stable Diffusion [46]



A polar bear is playing with a blue bottle underwater.

# Content

---

**Background/ 04**

**Author/ 19**

**Method/ 25**

**Experiment/ 38**

---



## First Author

### XuDong Frank Wang (王旭东)

Hi! I am a Ph.D. candidate in Berkeley AI Research (BAIR) lab at [UC Berkeley](#), advised by Prof. [Trevor Darrell](#). Prior to BAIR, I was a member of International Computer Science Institute (ICSI), advised by Prof. [Stella X. Yu](#). I received my Master's Degree (thesis plan) in Intelligent Systems, Robotics and Control at [University of California, San Diego](#), advised by Prof. [Nuno Vasconcelos](#) and Bachelor's Degree as an honorary member of Tang Aoqing Honors Program in Science at Jilin University.



[Google Scholar](#) / [Github](#) / [Twitter](#)

I am currently a student researcher at [Google DeepMind](#), working with Prof. [Cordelia Schmid](#), Dr. [Xingyi Zhou](#), and Dr. [Alireza Fathi](#). Previously, I was a research scientist intern in the [Generative AI Research \(GenAI\) team at Meta AI](#) from May 2023 to Dec 2023, and before that, in [Fundamental AI Research \(FAIR\) labs at Meta AI](#) from May 2022 to Dec 2022. During my tenure at Meta, I had the privilege of collaborating with Dr. [Ishan Misra](#) and Dr. [Rohit Girdhar](#). I also worked closely with Dr. [Zhaowei Cai \(Amazon AWS AI Labs\)](#), Dr. [Zhirong Wu \(Microsoft Research\)](#) and Prof. [Ziwei Liu \(NTU\)](#) from 2019 to 2022. I was a staff researcher at the vision group of [ICSI](#) between 2019 and 2020.

My research has been primarily focused on: [1] representation learning without reference to human annotations or under minimal human supervision. [2] text-to-image diffusion models. [3] grounded Large Multimodal Models (LMMs).

*How to pronounce XuDong? It is composed of two syllables: "Xu" and "Dong": /jü/-/do:ŋ/*

Contact: [xdwang \[at\] eecs \[dot\] berkeley \[dot\] edu](mailto:xdwang@eecs.berkeley.edu)

---

## 20 Author

---

Second Author



***Prof. Trevor Darrell***

**CS Division, University of California, Berkeley**

**Founding Co-Director, [Berkeley Artificial Intelligence Research \(BAIR\)](#), [Berkeley Deep Drive \(BDD\)](#), and [BAIR Commons](#).**

Prof. Darrell is on the faculty of the CS and EE Divisions of the EECS Department at UC Berkeley.

He received the S.M., and PhD. degrees from MIT in 1992 and 1996, respectively, and obtained his B.S.E. degree from the University of Pennsylvania in 1988.

---

# 21 Author

---

Third Author

## Saketh Rambhatla

I am a postdoctoral researcher at Meta AI working with Dr. [Ishan Misra](#). I obtained my Ph.D. in ECE at University of Maryland ([UMD](#)), College Park, where I worked on in-the-wild visual understanding with Dr. [Rama Chellappa](#) and Dr. [Abhinav Shrivastava](#).

I completed my bachelor's and master's at [Indian Institute of Technology, Kharagpur](#). During my master's I worked on SLAM and pose recognition.

[Email](#) / [CV](#) / [Google Scholar](#)





---

## 22 Author

---

Fourth Author



Rohit Girdhar

Research Scientist

GenAI Research, Meta

I am a Research Scientist in the GenAI Research group at Meta. My current research focuses on understanding and generating multimodal data, using minimal human supervision. I obtained a MS and PhD in Robotics from Carnegie Mellon University (here's a link to my [dissertation](#)), where I worked on learning from and understanding videos. I was previously part of the Facebook AI Research (FAIR) group at Meta, and have spent time at DeepMind, Adobe and Facebook as an intern. See [here](#) for a formal bio.

---

# 23 Author

---

## Fifth Author

### Ishan Misra


Research Scientist @ GenAI (Meta)

I work on computer vision and machine learning research specifically in generative AI and self-supervised learning. I am a Research Scientist in the GenAI group at Meta where I lead the research efforts on video generation models. I was the tech lead for [Meta's Movie Gen project](#) which released foundation models for video generation, video editing, video personalization, and audio generation.

Previously, I was part of the FAIR team at Meta where I worked on self-supervised learning in computer vision and multimodal learning.

I got my PhD at Carnegie Mellon University. I received [CMU's Recent Alumni Achievement Award](#) in 2024 for my research contributions to computer vision and machine learning.



 [ishanmisra@gmail.com](mailto:ishanmisra@gmail.com)

 [imisra@meta.com](mailto:imisra@meta.com)

 [@imisra\\_](#)

 [@imisra](#)

 [Google Scholar](#)



# Content

---

**Background/ 04**

**Author/ 10**

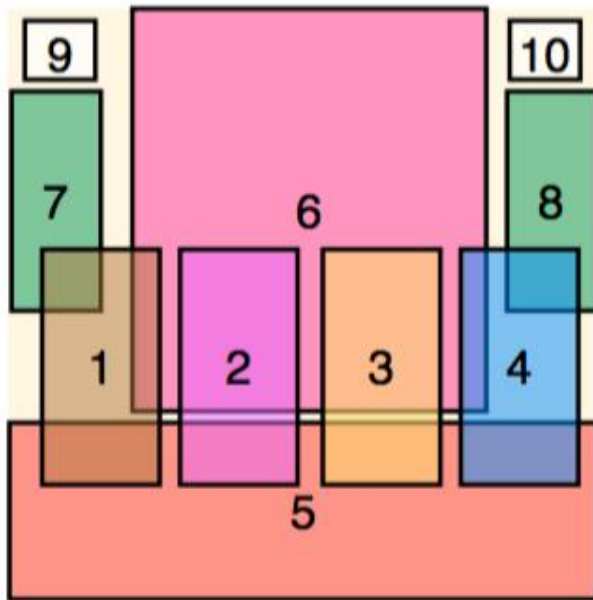
**Method/ 15**

**Experiment/ 40**

---

## 25 Method—functionalities

### Diverse Instance Attributes and Locations



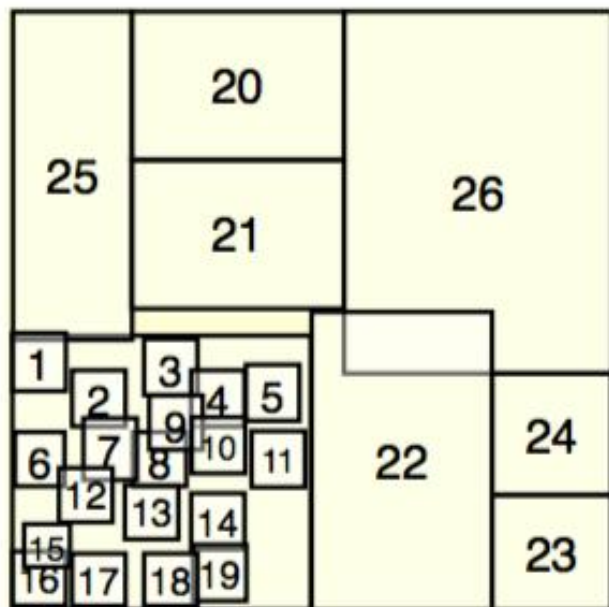
**Image Caption:** An image depicting in the morning. A **brown** cute teddy bear, a **purple** cute teddy bear, a **yellow** cute teddy bear, a **blue** cute teddy bear all standing side by side on a **red** brick road. The scene should be set in front of **Pink** Castle with clear **blue** sky overhead, punctuated by fluffy **white** clouds, and trees with **green** leaves. The **pink** castle should loom majestically in the background. **Instance Captions:** 1-4) A **brown/purple/yellow/blue** teddy bear; 5) a **red** brick road; 6) **Pink** Castle; 7-8) **green** leaves; 9-10) fluffy **white** clouds

---

## 26 Method—functionalities

---

Dense small objects



**Image Caption:** Craft an oil painting: Picture a seaside garden drenched in radiant hues of roses, lilies, and lavender, transitioning gracefully into the expansive azure ocean and blue sky. Integrate a weathered, rustic pathway with steps that invite viewers towards the water's edge, complemented by a prominent bouquet of flowers and plants.

**Instance Captions:** 1-19) roses; 20) sky; 21) ocean; 22) pathway with steps; 23) bouquet of flowers; 24) plant; 25-26) plants

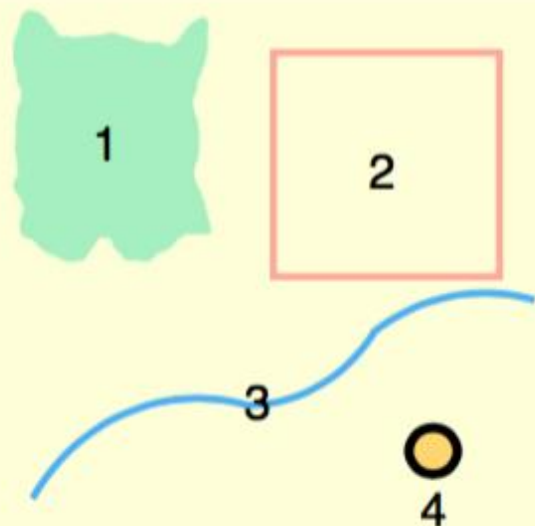


---

## 27 Method—functionalities

---

Various Location Conditions (box, mask, scribble, point)



**Image Caption:** An image of two little husky puppy in a wicker basket.

**Instance Captions:** 1) a husky puppy sitting in a wicker basket + *Mask*. 2) a black and white husky puppy in a blue towel + *Box*. 3) two husky puppies sitting in a wicker basket + *Scribble*. 4) a blue towel + *Point*

---

## 28 Method—contributions

---

### Contributions:

- propose and study instance-conditioned image generation that allows flexible location and attribute specification for multiple instances.
- propose **UniFusion** which projects various forms of instance level conditions into the same feature space to enable multiple control formats.
- propose **ScaleU** to further improve generation quality.
- propose **Multi-instance Sampler**, which reduces information leakage between multiple instances.
- a **dataset** with instance-level captions and a new set of **evaluation benchmarks** and **metrics** for measuring the performance of location grounded image generation.



## 29 Method—problem definition



**Image Caption:** An image of two little husky puppy in a wicker basket.

**Instance Captions:** 1) a husky puppy sitting in a wicker basket + *Mask*. 2) a black and white husky puppy in a blue towel + *Box*. 3) two husky puppies sitting in a wicker basket + *Scribble*. 4) a blue towel + *Point*

Image generation model:  $f(\mathbf{c}_g, \{(\mathbf{c}_1, \mathbf{l}_1), \dots, (\mathbf{c}_n, \mathbf{l}_n)\})$

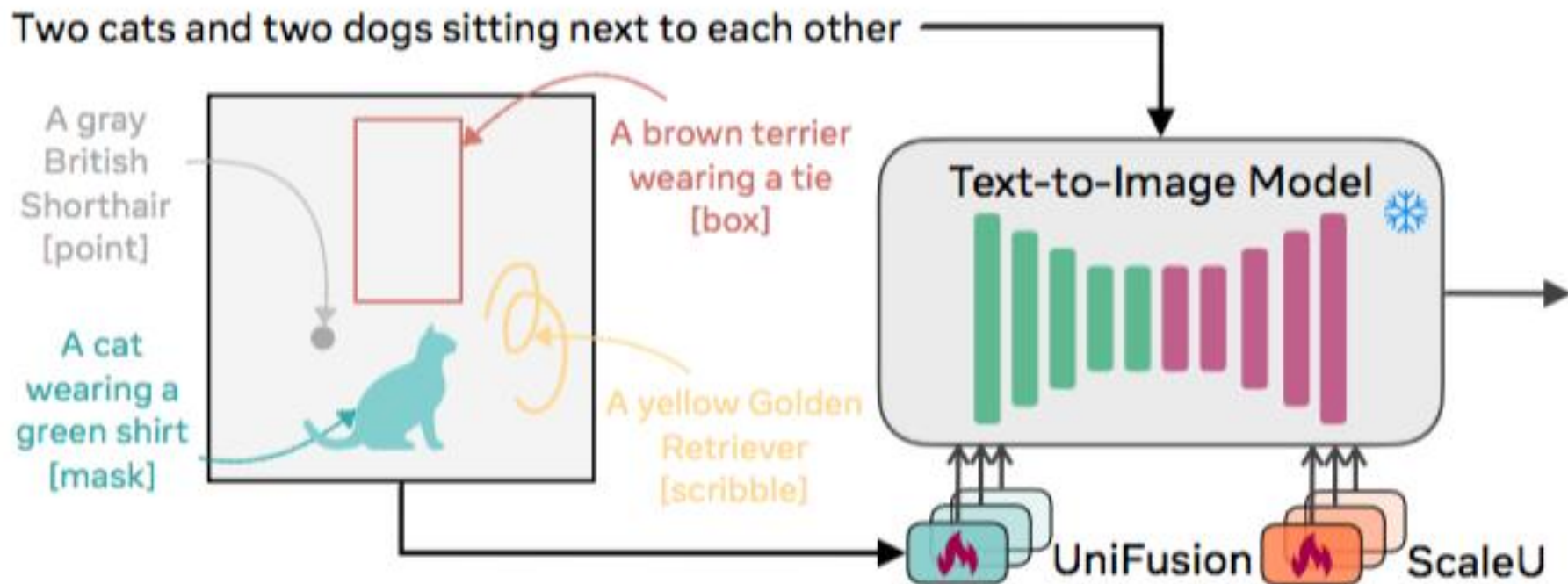
$\mathbf{c}_g$

Global text caption

$(\mathbf{c}_i, \mathbf{l}_i)$

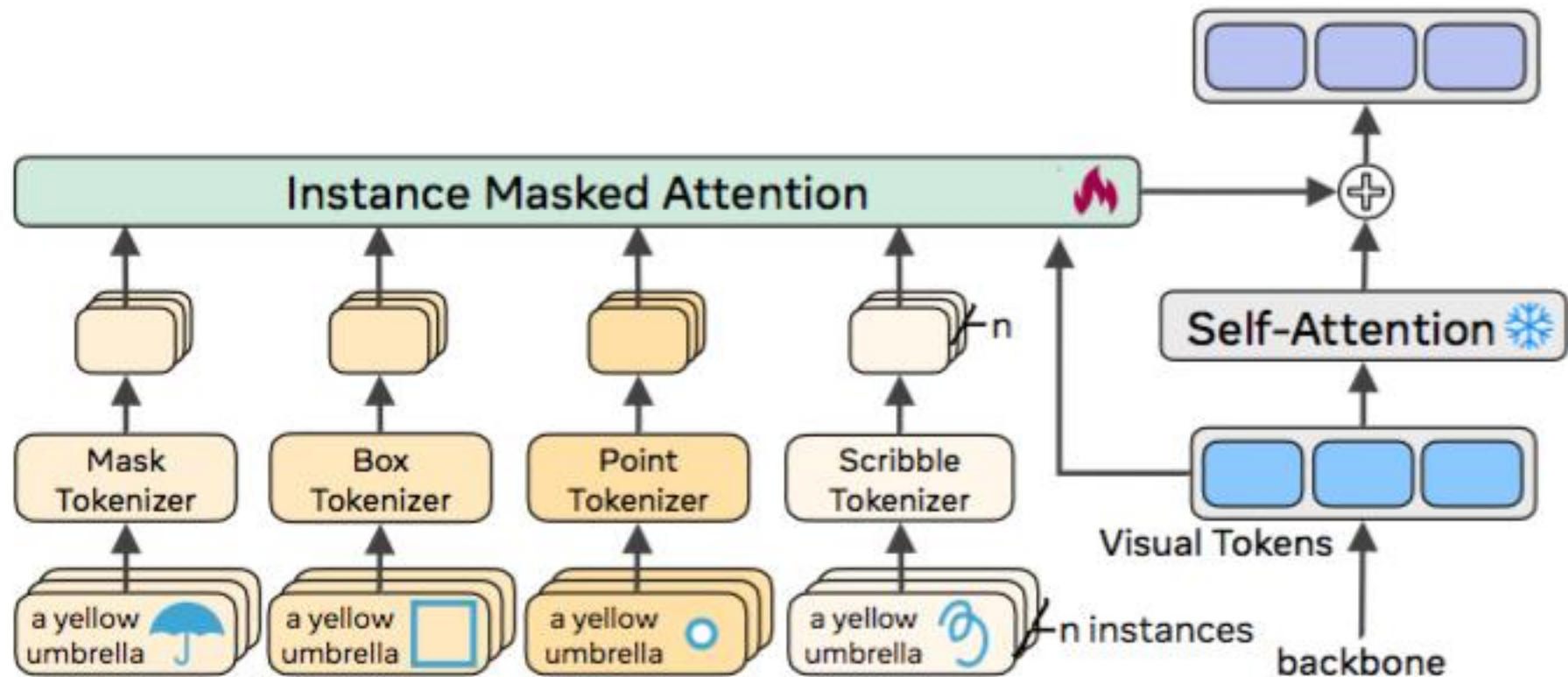
per-instance conditions containing caption  $c_i$  and location  $l_i$  for  $n$  instances

## 30 Method—UniFusion



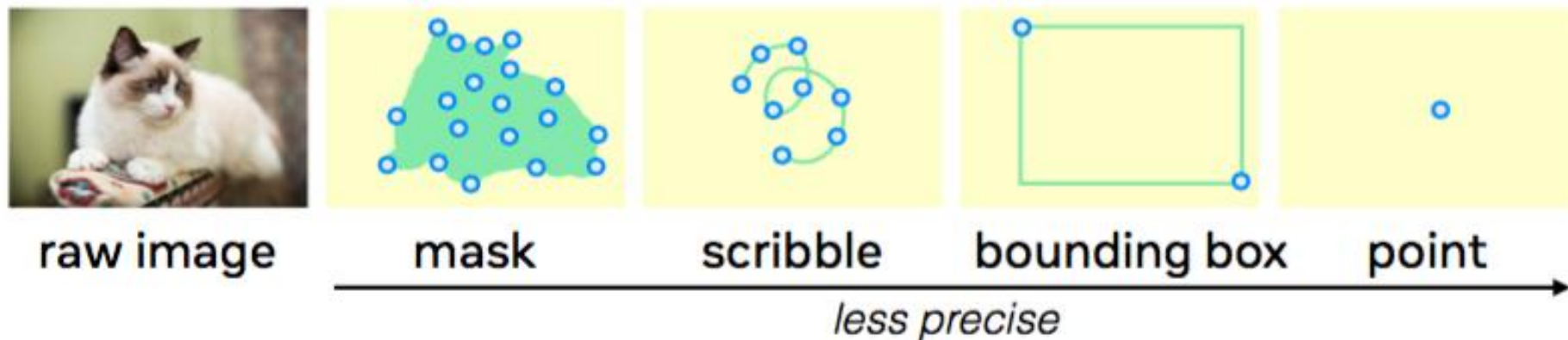
In addition to a global text prompt, InstanceDiffusion allows for paired instance-level prompts and their locations to be specified when generating images.

## 31 Method—UniFusion



UniFusion projects various forms of instance-level conditions into the same feature space.

## 32 Method—UniFusion



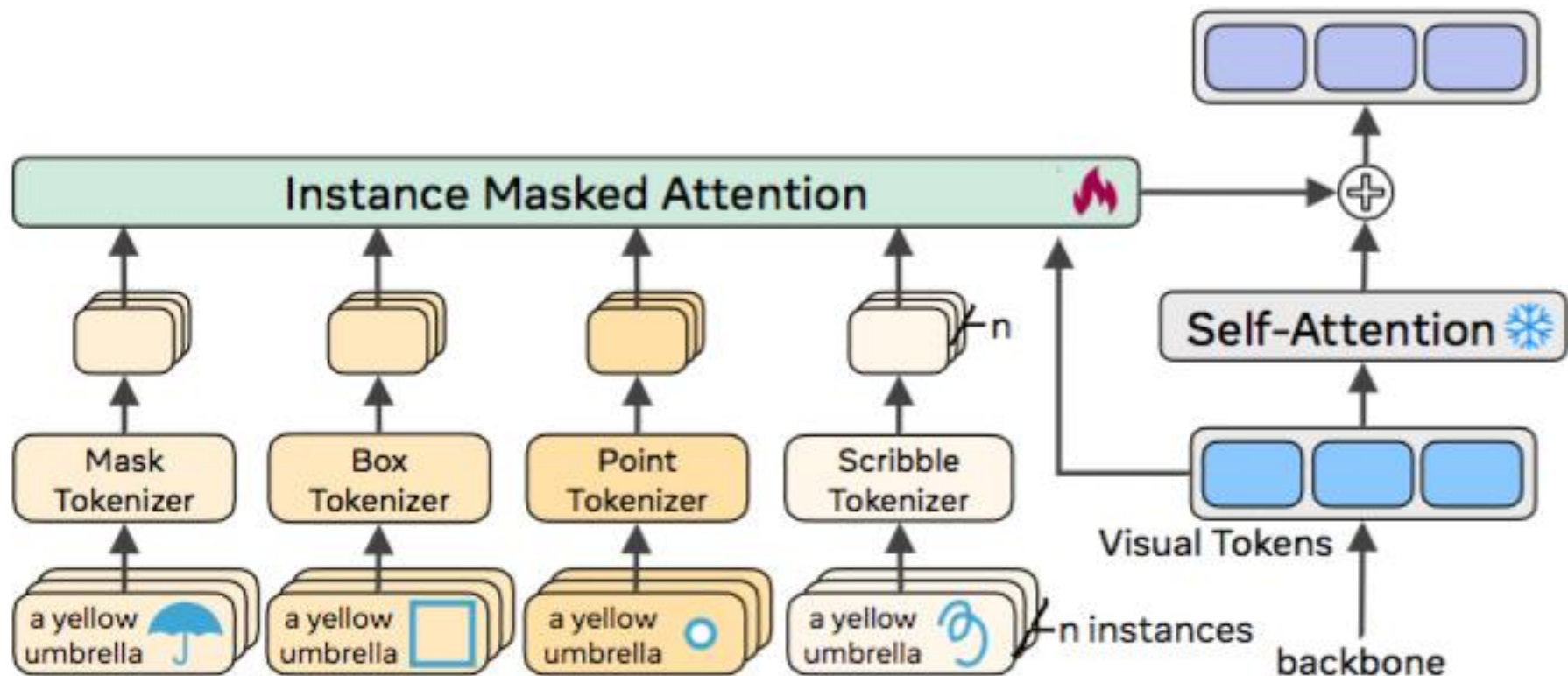
Convert the four location formats—masks, boxes, scribbles, single point—into 2D points:

$$\mathbf{p}_i = \{(x_k, y_k)\}_{k=1}^n$$

- A scribble is converted into a set of uniformly sampled points along the curve.
- A bounding box is parameterized by its top-left and bottom-right corners.
- For instance masks, we convert them into a set of points sampled from within the mask and from boundary polygons.



## 33 Method—UniFusion

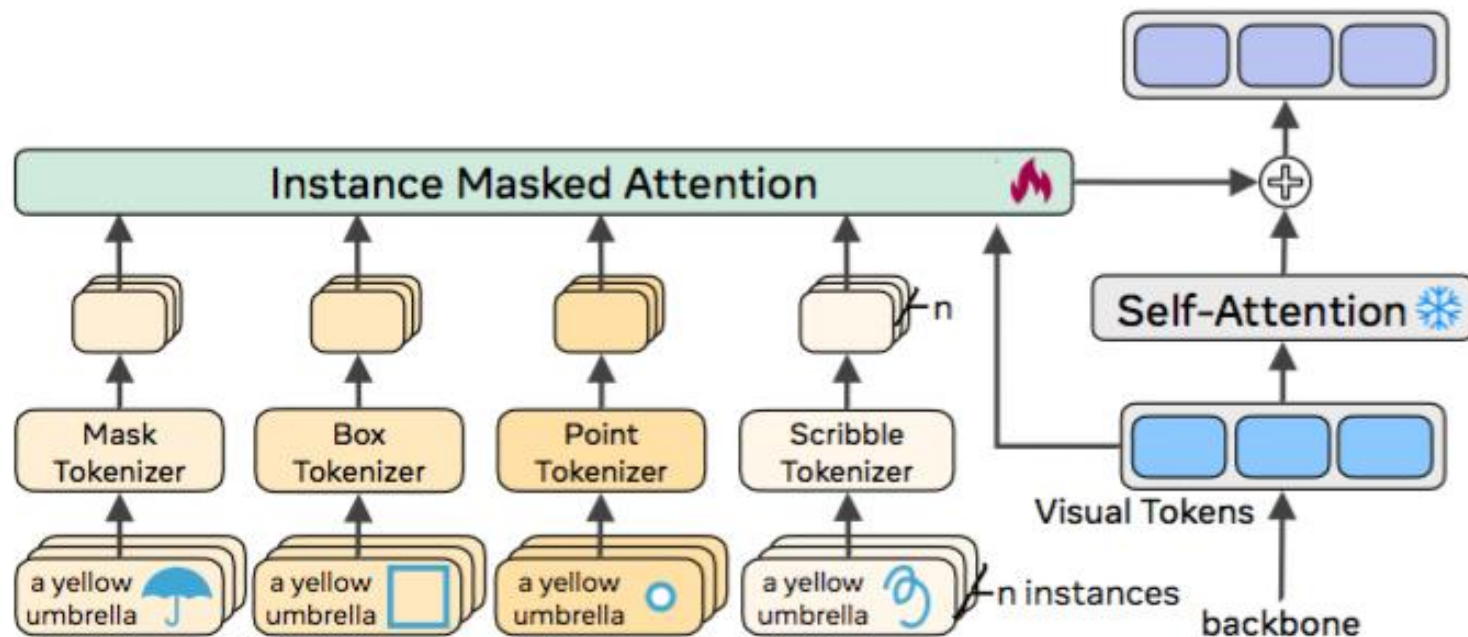


$$\mathbf{g}_i = \text{MLP}([\tau_\theta(\mathbf{c}_i), \gamma(\mathbf{p}_i)])$$

CLIP text encoder

Fourier mapping

## 34 Method—UniFusion

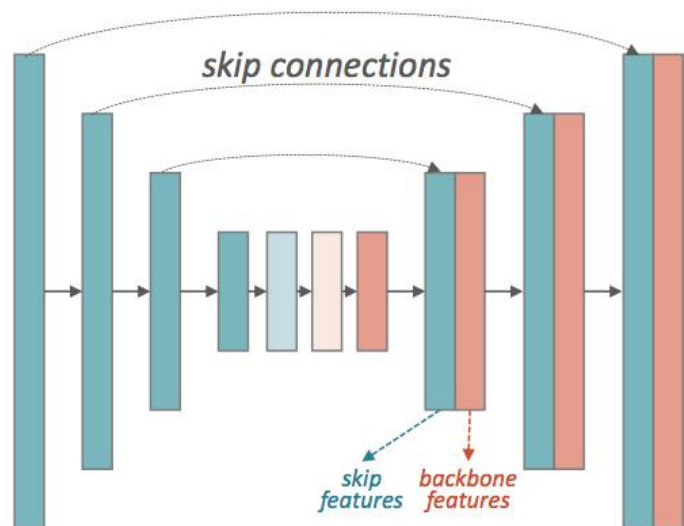


$$\tilde{\mathbf{V}} = \text{SA}_{\text{mask}}([\mathbf{V}, \mathbf{G}^{\text{mask}}, \mathbf{G}^{\text{scribble}}, \mathbf{G}^{\text{box}}, \mathbf{G}^{\text{point}}])$$

$$\text{mask for } \mathbf{v}_k \cdot \mathbf{v}_j^T : \mathbf{M}_{k,j} = -\text{inf if } I_{\mathbf{v}_k} \neq I_{\mathbf{v}_j}$$

$$\text{mask for } \mathbf{v}_k \cdot \mathbf{g}_i^T : \mathbf{M}_{k,m+i} = -\text{inf if } I_{\mathbf{v}_k} \neq i$$

## 35 Method—ScaleU



$\mathbf{F}_s$  Skip features

$\mathbf{F}_b$  Backbone features

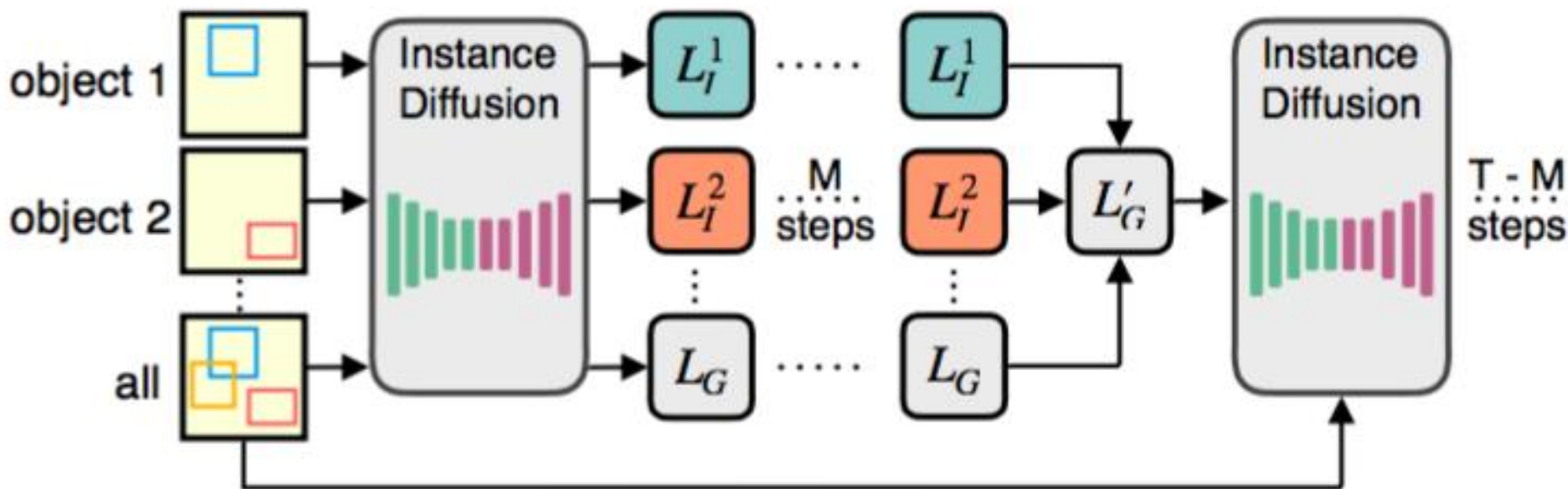
$$F'_b = F_b \otimes (\tanh(S_b) + 1)$$

$$F'_s = IFFT(FFT(F_s) \otimes \alpha)$$

$$\alpha(r) = \tanh(S_s) + 1 \text{ if } r < r_{thresh} \text{ otherwise } 1$$

$$\min_{\theta'} \mathcal{L}_{\text{Grounding}} = \mathbb{E}_{\mathbf{z}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} [\|\epsilon - f_{\{\theta, \theta'\}}(\mathbf{z}_t, t, \mathbf{y})\|_2^2].$$

## 36 Method—Multi-instance Sampler



To further minimize the information leakage across multiple instance conditionings, Multi-instance Sampler is proposed and adopted during model inference which improves the quality and fidelity of the generated image.



# Content

---

**Background/ 04**

**Author/ 19**

**Method/ 25**

**Experiment/ 38**

---

---

## 38 Experiment—data collection

---

Construct instance-level annotated training dataset:

**1. Image-level label generation:**

employ RAM model to generate a list of image-level tags as global caption

**2. Bounding box and mask generation:**

use Grounded-SAM to produce instance-level bounding boxes and masks

**3. Instance-level text prompt generation:**

crop the instances using the corresponding bounding boxes and feed to the BLIP-V2 model to produce instance-level text prompts.

---

## 39 Experiment—data collection

---

### **4. Obtain scribbles:**

obtain scribble by randomly sampling points within the masks and connecting them with bezier curve.

### **5. Obtain single-points:**

randomly select a point within a circular region of radius  $0.1 \cdot r$ , centered at the bounding box's center, where  $r$  is the length of the shortest side of the box.

---

## 40 Experiment—data collection

---

### Test data collection

- COCO val with 80 classes
- LVIS val with over 1200 classes

Do not use the real images from the dataset, and only use the text and location conditions.



---

## 41 Experiment—evaluation metrics

---

Evaluation metrics for alignment to instance locations:

### **Bounding Box**

Use the pre-trained YOLOv8-Det detection model. Compare the model's detected bounding boxes on the generated image with the bounding boxes specified in the input using COCO's official evaluation metrics (AP and AR).

---

## 42 Experiment—evaluation metrics

---

Evaluation metrics for alignment to instance locations:

### **Instance Mask**

Use the pre-trained YOLOv8-Seg segmentation model to detect instance masks in the generated image, and compare the detected masks with the specified masks in the input using the COCO AO and AR metrics and IOU score.

---

## 43 Experiment—evaluation metrics

---

Evaluation metrics for alignment to instance locations:

### **Scribble**

Introduced a new evaluation metric using YOLOv8-Seg. We report “Points in Mask” (PiM), which measures how many of randomly sampled points in the input scribble lie within the detected mask.

---

## 44 Experiment—evaluation metrics

---

Evaluation metrics for alignment to instance locations:

### **Single point**

Similar to scribble, the instance-level accuracy PiM is 1 if the input point is within the detected mask, and 0 otherwise. We then calculate the averaged PiM score.



---

## 45 Experiment—evaluation metrics

---

Evaluation metrics for alignment to instance locations:

### **Compositional attribute binding**

To measure if the generated instances adhere to the attribute (color and texture) specified in the instance prompts. We use YOLOv8-Det to detect the bounding boxes. We feed the cropped box to the CLIP model to predict its attribute (colors and textures), and measure the accuracy of the prediction with respect to the attribute specified in the instance prompt.

We use 8 common colors, *i.e.*, “black”, “white”, “red”, “green”, “yellow”, “blue”, “pink”, “purple”, and 8 common textures, *i.e.*, “rubber”, “fluffy”, “metallic”, “wooden”, “plastic”, “fabric”, “leather” and “glass”.

---

## 46 Experiment—evaluation metrics

---

Evaluation metrics for alignment to instance locations:

### **Instance text-to-image alignment**

Report the CLIP-Score on cropped object to measure the distance between the instance text prompt's features and the cropped object images.

---

## 47 Experiment—evaluation metrics

---

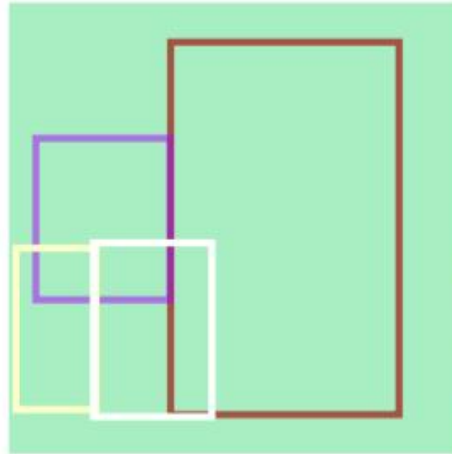
Evaluation metrics for alignment to instance locations:

### **Global text-to-image alignment**

Report CLIP-Score between the input text prompt and the generated image.

# 48 Experiment—qualitative results

InstanceDiffusion



**Image Caption:**  
Little Terrier Puppy with a bouquet of flowers on a blurred **green background**

**Instance Captions:**

- **purple** flowers
- **yellow** flowers
- **white** flowers
- **a black and tan** yorkshire terrier puppy

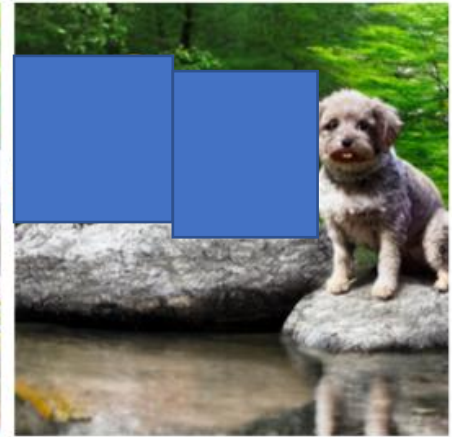
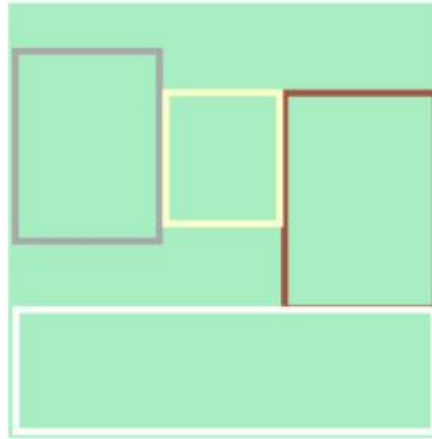
GLIGEN





# 49 Experiment—qualitative results

InstanceDiffusion



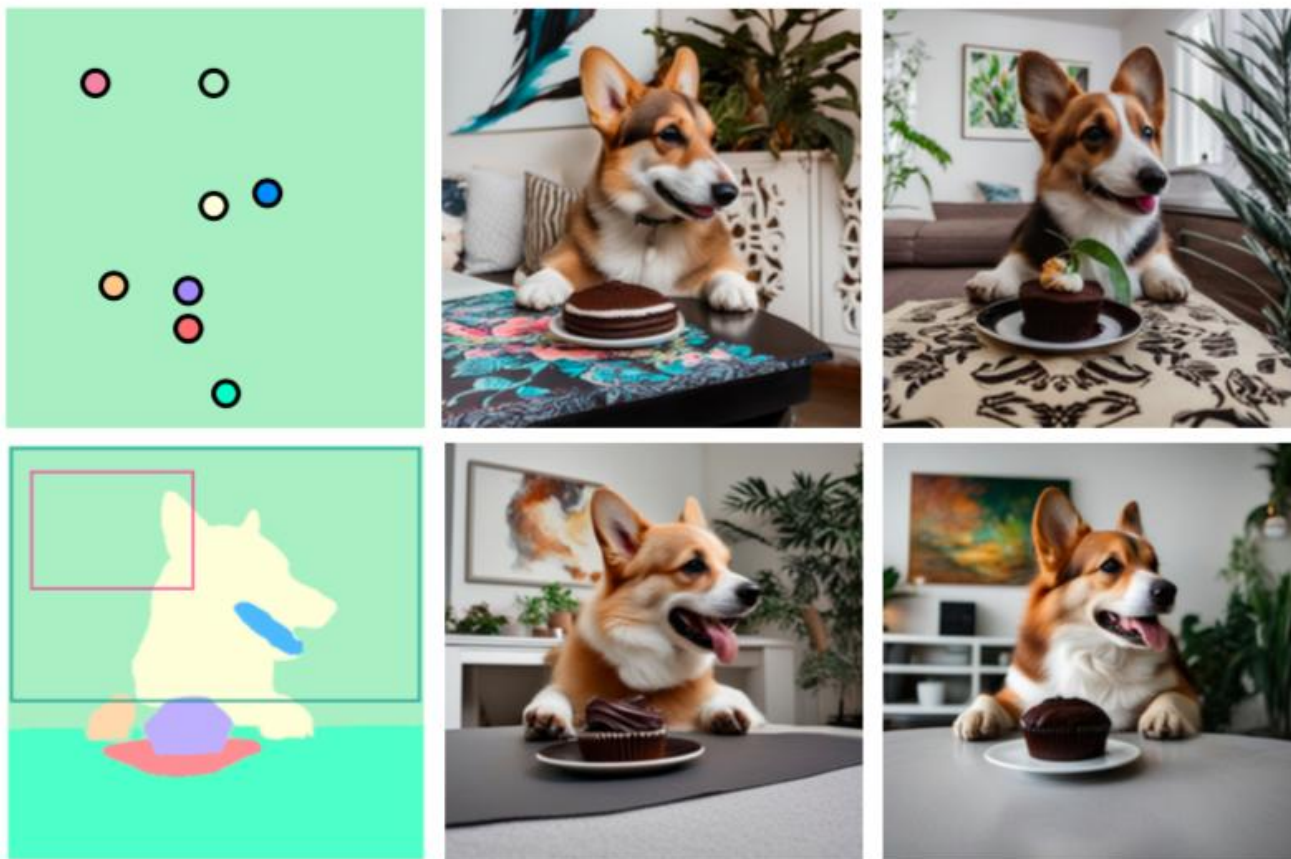
**Image Caption:**  
a yellow American robin,  
brown Maltipoo dog, a gray  
British Shorthair in a stream,  
alongside with trees and rocks

- Instance Captions:**
- a gray British Shorthair
  - a yellow American robin
  - a brown Maltipoo dog
  - a close up of a small waterfall in the woods

GLIGEN

## 50 Experiment—qualitative results

Results of multiple instance control formats



**Image Caption:** Cute Corgi at table in a living room with plants and painting on the wall. A chocolate cake is on the table. **Instance Captions:** 1) a Corgi sitting in front of a cupcake 2) Corgi's mouth and tongue 3) a plate 4) a chocolate cupcake on a plate 5) a white paw 6) a table 7) a living room with plants 8) oil painting on the wall

# 51 Experiment—quantitative results

Leading performance in all the metrics

Location format input → Method	Boxes				IoU	Instance Masks				Points		Scribble	
	AP <sup>box</sup>	AP <sub>50</sub> <sup>box</sup>	AR <sup>box</sup>	FID (↓)		AP <sup>mask</sup>	AP <sub>50</sub> <sup>mask</sup>	AR <sup>mask</sup>	FID (↓)	PiM	FID (↓)	PiM	FID (↓)
Upper bound (real images)	50.2	66.7	61.0	-	-	40.8	63.5	58.0	-	-	-	-	-
GLIGEN [34]	19.6	35.0	30.7	27.0	-	-	-	-	-	-	-	30.2 <sup>†</sup>	32.4 <sup>†</sup>
ControlNet [65] <sup>‡</sup>	-	-	-	-	-	6.5	13.8	12.9	-	-	-	-	-
DenseDiffusion [28]	-	-	-	-	35.0 / 48.6	-	-	-	-	-	-	-	-
SpaText [4] <sup>‡</sup>	-	-	-	-	-	5.3	12.1	10.7	-	-	-	-	-
<b>InstanceDiffusion</b>	<b>38.8</b>	<b>55.4</b>	<b>52.9</b>	<b>23.9</b>	<b>61.6 / 71.4</b>	<b>27.1</b>	<b>50.0</b>	<b>38.1</b>	<b>25.5</b>	<b>81.1</b>	<b>27.5</b>	<b>72.4</b>	<b>27.3</b>
<i>vs. prev. SoTA</i>	<b>+19.2</b>	<b>+20.4</b>	<b>+21.8</b>	<b>-3.1</b>	<b>+25.4 / +22.8</b>	<b>+20.6</b>	<b>+36.2</b>	<b>+25.2</b>	-	-	-	<b>+42.2</b>	<b>-4.9</b>
InstanceDiffusion (hybrid)	44.6	59.6	58.8	25.5	-	-	-	-	-	86.0	25.5	82.9	26.4

**Table 1. Evaluating different location formats** as input when generating images. We measure the YOLO recognition performance (AP, AR) for the generated image wrt the location condition provided as inputs, and FID on the COCO val set. Most prior methods only support a handful of the location conditions. We observe that InstanceDiffusion, while using the same model parameters, supports various location inputs. In each setting, InstanceDiffusion substantially outperforms prior work on all metrics. <sup>†</sup>: GLIGEN’s scribble-based results are derived by using the top-right and bottom-left corners as the bounding box for the region encompassed by the scribble. We measure the IoU using [28]’s official evaluation codes (left), and YOLOv8-Seg (right). <sup>‡</sup>: ControlNet [65] (and SpaText [4]) only supports *semantic* segmentation mask inputs, and do not differentiate between instances of the same class. We assess ControlNet’s AP<sup>mask</sup> using its official mask conditioned Image2Image generation pipeline. Hybrid: we add instance masks as additional conditions.



## 52 Experiment—quantitative results

### More quantitative comparison with GLIGEN

Methods	Color		Texture		Human Eval
	Acc <sup>color</sup>	CLIP <sup>local</sup>	Acc <sup>texture</sup>	CLIP <sup>local</sup>	
GLIGEN	19.2	0.206	16.6	0.206	19.7
InstDiff	<b>54.4</b>	<b>0.250</b>	<b>26.8</b>	<b>0.225</b>	80.3
$\Delta$	<b>+35.2</b>	<b>+0.044</b>	<b>+10.2</b>	<b>+0.019</b>	

**Table 2. Attribute binding.** We measure whether the attributes of the generated instances match the attributes specified in the instance captions. We observe that InstanceDiffusion outperforms prior work on both types of attributes. Human evaluators prefer our generations significantly more than the prior work.

Methods	AP	AP <sub>50</sub>	AP <sub>s</sub>	AP <sub>m</sub>	AP <sub>l</sub>	AP <sub>r</sub>	AP <sub>c</sub>	AP <sub>f</sub>
Upper bound	44.6	57.7	33.2	55.0	66.1	31.4	44.5	50.5
GLIGEN [34] <sup>†</sup>	9.9	9.5	1.6	10.5	31.1	7.4	10.0	10.9
InstanceDiffusion	17.9	25.5	5.5	24.2	45.0	12.7	18.7	19.3
<i>vs. prev. SoTA</i>	<b>+8.0</b>	<b>+16.0</b>	<b>+3.9</b>	<b>13.7</b>	<b>+13.9</b>	<b>+5.3</b>	<b>+8.7</b>	<b>+8.4</b>

**Table 3. Box inputs on LVIS val.** We evaluate using a pretrained detector (ViTDet-L [33]) and obtain the upper bound by evaluating the detector on real images resized to 512×512. InstanceDiffusion significantly outperforms prior work across all metrics including object sizes, and class frequencies. <sup>†</sup>: reproduced results.

## 53 Experiment—ablation study

### Ablation study results

#	FA Fusion	MaskAttn	ScaleU	Inst. Cap.	MIS	AP <sub>50</sub> <sup>mask</sup>	Acc <sup>color</sup>	FID (↓)
1	✓	✓	✓	✓	✓	50.0	55.4	25.5
2	✗	✓	✓	✓	✓	45.5(5.5)	49.4(6.0)	25.8(0.3)
3	✓	✗	✓	✓	✓	49.3(0.7)	53.1(2.3)	25.7(0.2)
4	✓	✓	✗	✓	✓	47.7(2.3)	52.2(3.2)	25.7(0.2)
5	✓	✓	✓	✗	✓	47.8(2.2)	38.2(17.2)	25.6(0.1)
6	✓	✓	✓	✓	✗	49.8(0.2)	49.5(5.9)	28.6(3.1)

**Table 5. Contribution of each component** evaluated by removing or adding it and measuring the impact of the generated image in terms of its instance location performance (AP), and instance attribute binding (Acc), and overall image quality (FID). When Format Aware (FA) fusion mechanism is disabled, we use the Joint format fusion mechanism instead. Top row is the default setting for InstanceDiffusion in the paper and we report the drop in performance for each subsequent row in **red**.

% of Steps →	0%	10%	20%	30%	36%	40%	50%
FID	28.6	27.8	27.4	25.8	25.5	25.0	27.0
AP <sub>50</sub> <sup>mask</sup>	49.8	49.8	49.4	49.4	50.0	49.2	48.3

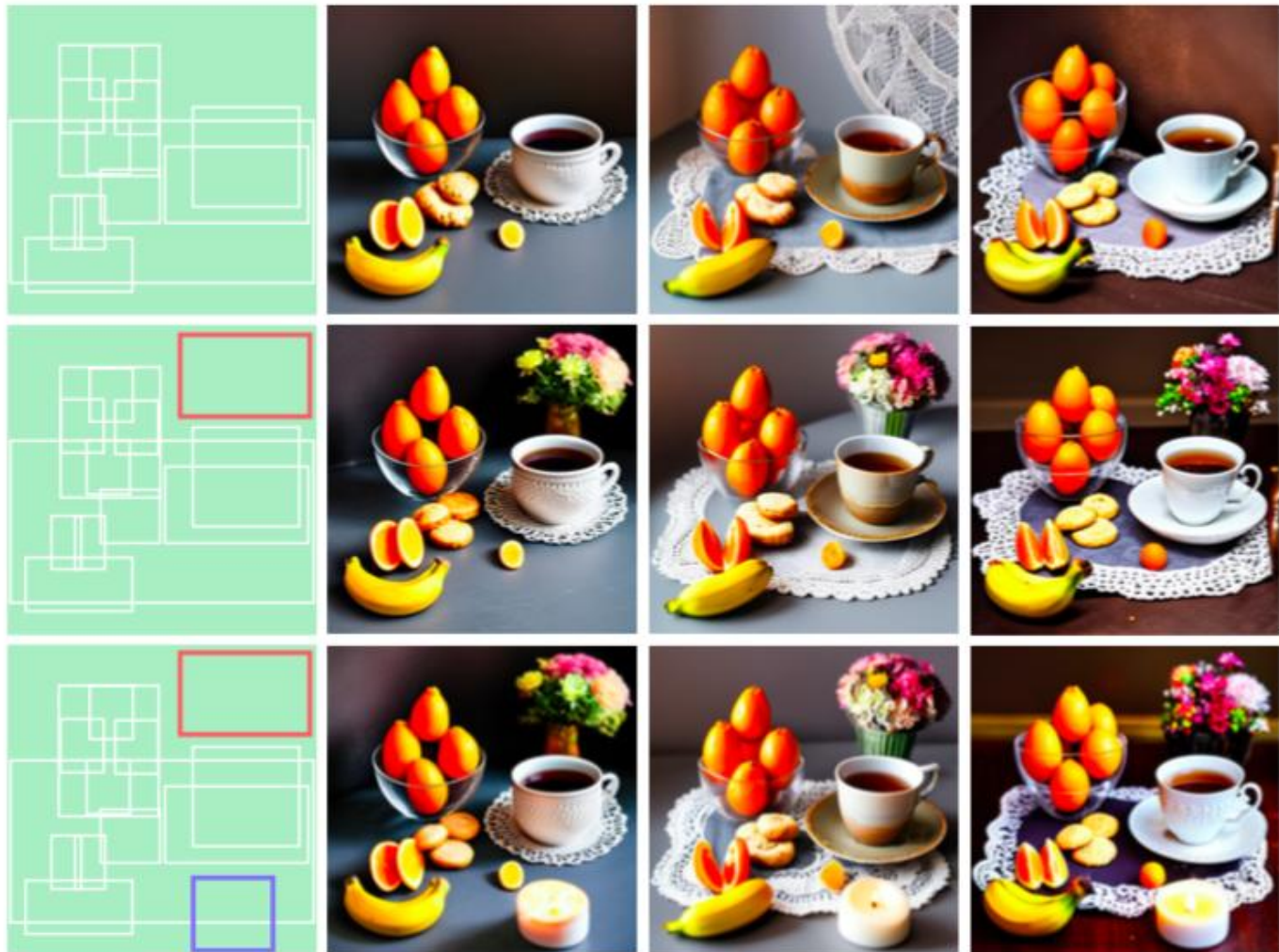
**Table 7. Multi-instance Sampler (MIS)** lowers the FID and improves overall image quality. Location conditions: instance masks.

	GLIGEN [34]	w/ MIS	InstanceDiffusion	w/ MIS
Acc <sup>color</sup>	19.2	<b>29.7</b>	49.5	<b>55.4</b>

**Table 8. Multi-instance Sampler** can be adapted for previous location conditioned work, yielding notable performance gains.



# 54 Experiment—application: iterative generation



**Image Caption:** A cup of tea with tangerines, bananas, and cookies on the table. high quality. professional photo.  
**Instance Captions:** 1) a cup of tea on a lace doily 2) a close up of three oranges on a black background 3) oranges in a glass bowl on a table 4) a tray of pastries on a table with oranges 5) a close up of some cookies on a table 6) oranges in a glass bowl 7) oranges in a glass bowl 8) an orange that has been cut in half on a table 9) an orange is cut in half 10) bananas 11) a bouquet of flowers on a table 12) a bouquet of flowers on a table 13) A candle

**Thanks!**