南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

InstanceDiffusion: Instance-level Control for Image Generation

Xudong Wang[1,2] Trevor Darrell[2] Sai Saketh Rambhatla[1] Rohit Girdhar[1] Ishan Misra[1]
[1]GenAI, Meta [2]UC Berkeley

project page: https://people.eecs.berkeley.edu/∼xdwang/projects/InstDiff/

https://github.com/frank-xwang/InstanceDiffusion

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

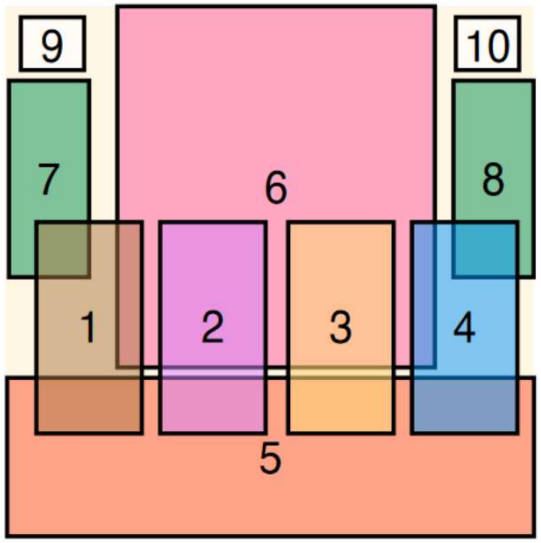


Image Caption: An image depicting in the morning. A ***brown*** cute teddy bear, a ***purple*** cute teddy bear, a ***yellow*** cute teddy bear, a ***blue*** cute teddy bear all standing side by side on a ***red*** brick road. The scene should be set in front of ***Pink*** Castle with clear ***blue*** 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

a) Diverse Instance Attributes and Locations

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

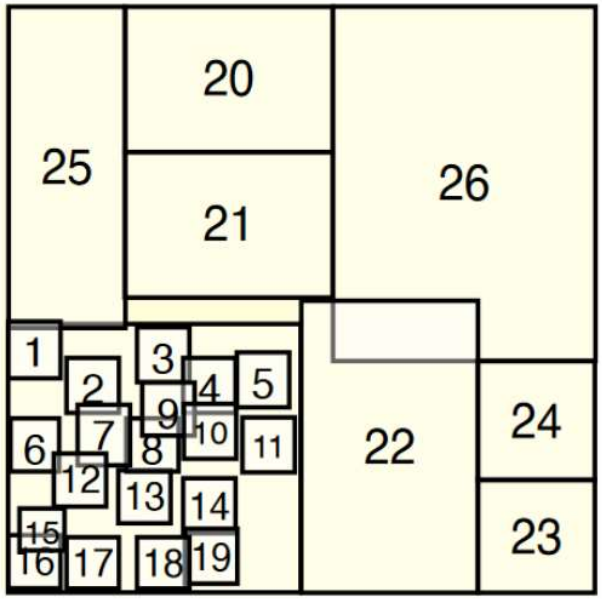


Image Caption: Craft an oil painting: Picture a seaside garden drenched in radiant hues of roses, lilies, and lavender, transitioning gracefully into the expansive azure ocean and blue sky. Integrate a weathered, rustic pathway with steps that invite viewers towards the water's edge, complemented by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

b) Dense Small Objects

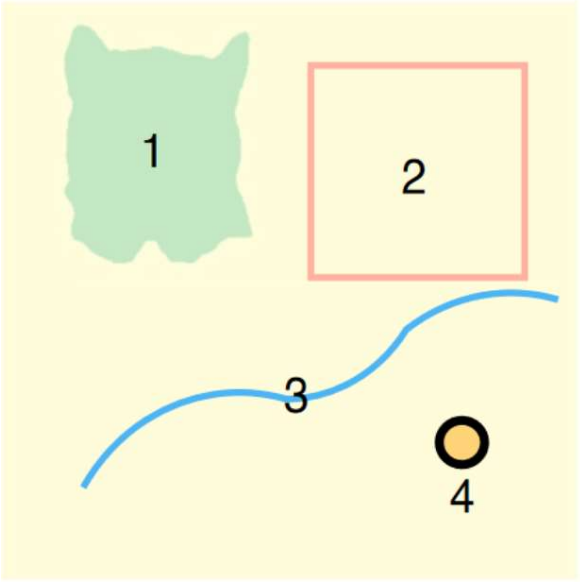


Image Caption: An image of two little husky puppy in a wicker basket.
Instance Captions: 1) a husky puppy sitting in a wicker basket + ***Mask.*** 2) a black and white husky puppy in a blue towel + ***Box***. 3) two husky puppies sitting in a wicker basket + ***Scribble***. 4) a blue towel + ***Point***

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

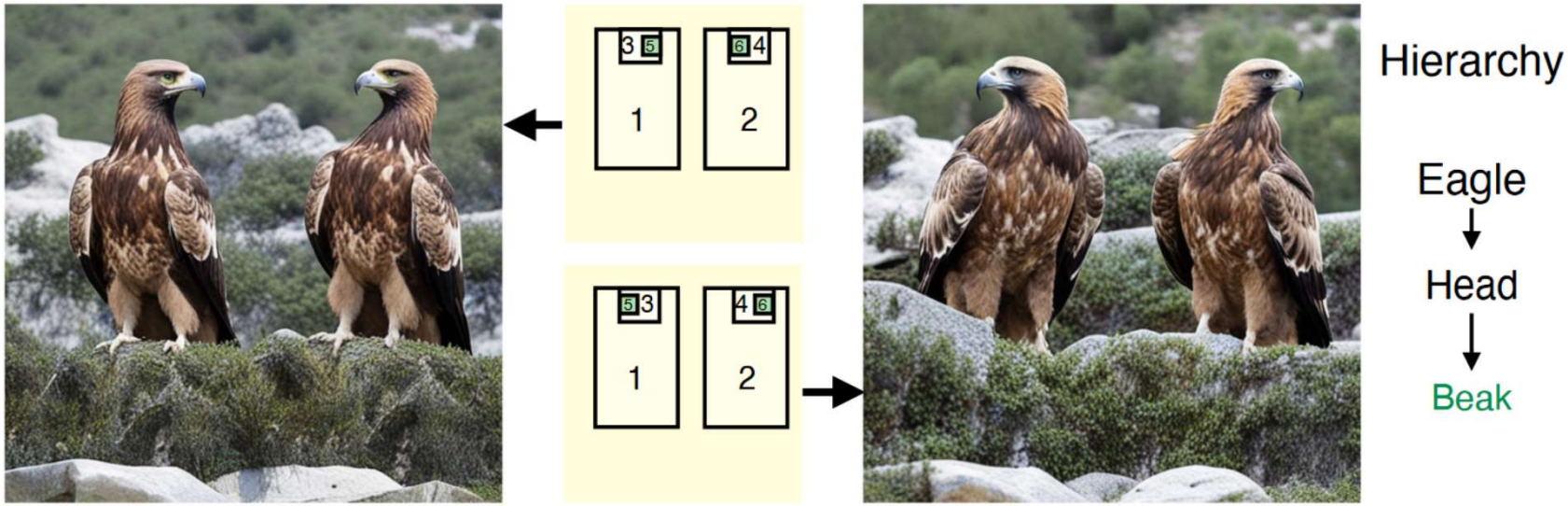
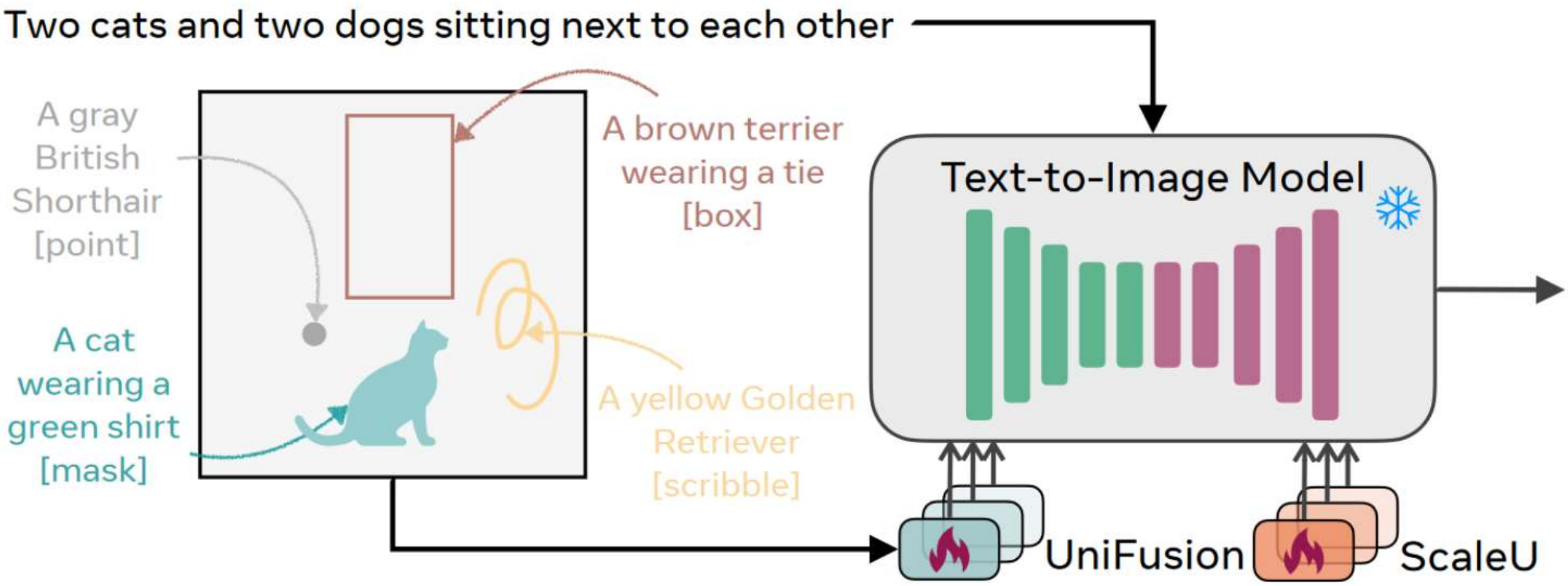


Image Caption: A golden eagle perched on a rugged rock.
Instance Caption: 1-2) A golden eagle; 3-4) Eagle's head; 5-6) **Eagle's beak**

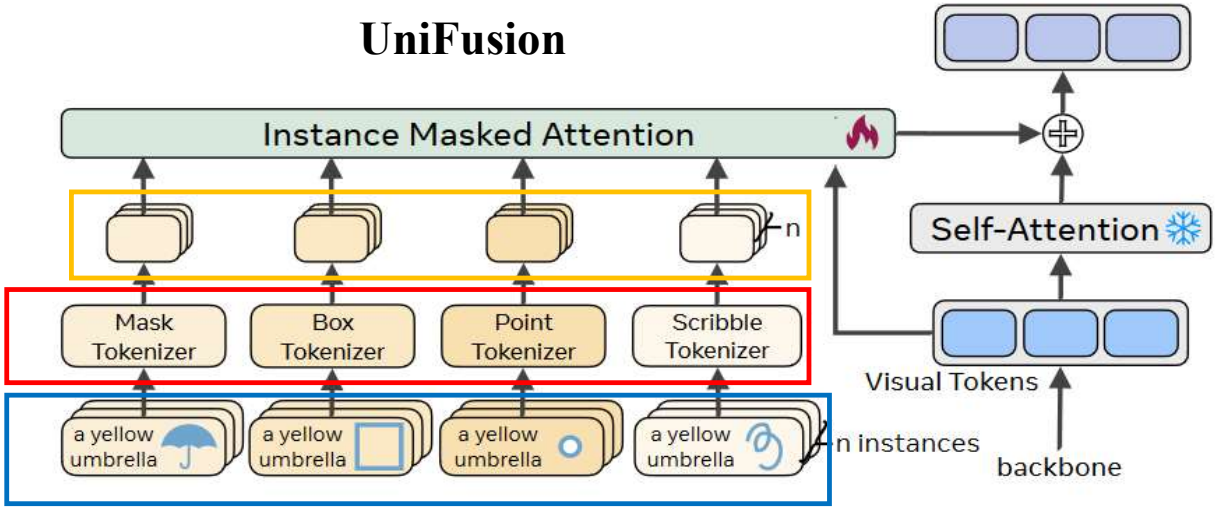
d) Image Composition with Whole Instance, Part and Subpart

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

InstanceDiffusion



UniFusion



Instance-Masked Attention and Fusion Mechanism

$\mathbf{G}^{\text{mask}}$ $\mathbf{G}^{\text{scribble}}$ $\mathbf{G}^{\text{box}}$ $\mathbf{G}^{\text{point}}$

$$\widetilde{\mathbf{V}} = \text{SA}_{\text{mask}}([\mathbf{V}, \mathbf{G}^{\text{mask}}, \mathbf{G}^{\text{scribble}}, \mathbf{G}^{\text{box}}, \mathbf{G}^{\text{point}}]) \quad (2)$$

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

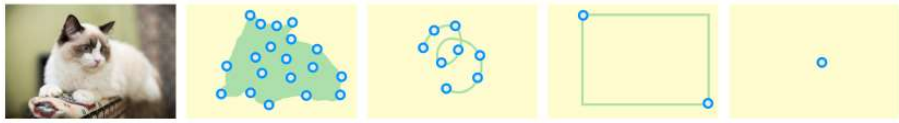
mask for $\mathbf{v}_k \cdot \mathbf{g}_i^T$: $\mathbf{M}_{k,m+i} = -\inf$ if $I_{\mathbf{v}_k} \neq i$

where $I_{\mathbf{v}_k} = i$ if the visual token $\mathbf{v}_k$ falls within the region of the instance i defined by either a bounding box or an instance segmentation mask.

$$\mathbf{V} = \mathbf{V} + \tanh(\omega)\widetilde{\mathbf{V}}[:m]$$

where ω is a learnable parameter, initialized to 0

Location parameterization



masks : sparsely sampled points within the mask and uniformly sampled points from boundary polygons

scribble : uniformly sampled points

bounding boxes : the top-right and bottom-right corners

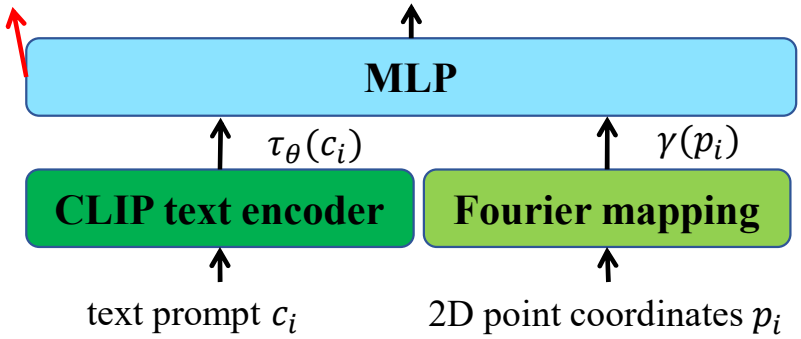
single point

location formats → 2D points $\mathbf{p}_i = \{(x_k, y_k)\}_{k=1}^{n}$

Instance Tokenizer

different for each location format

$g_i = MLP([\tau_\theta(c_i), \gamma(p_i)])$

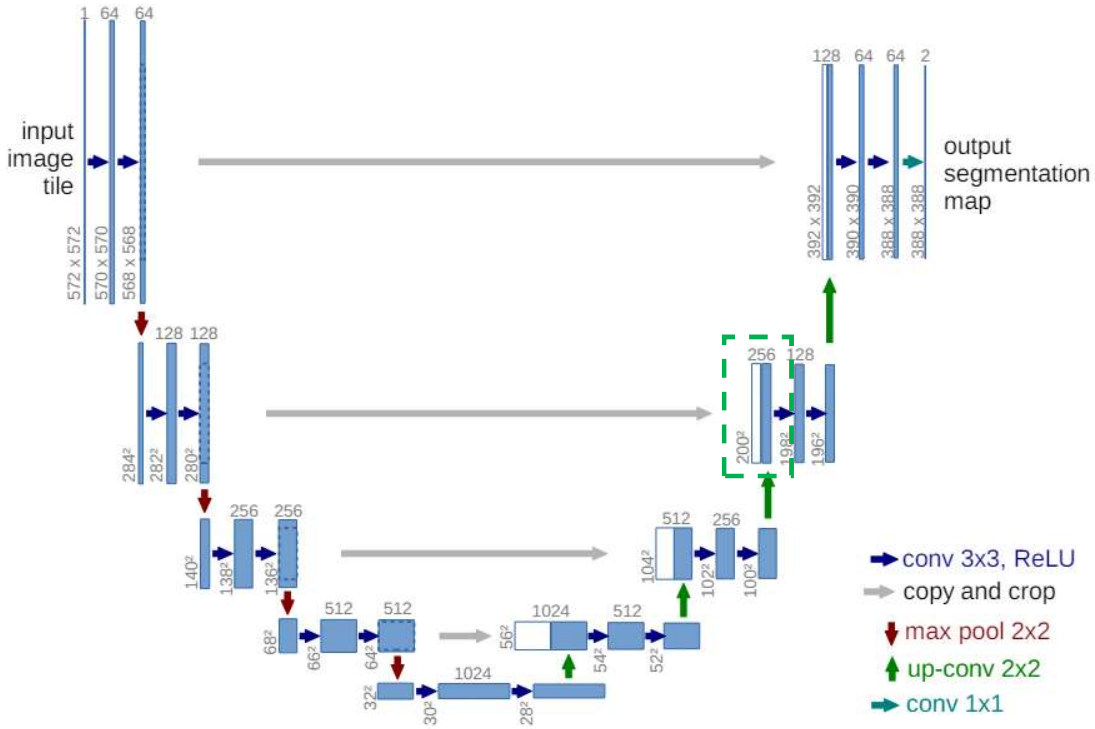


Final token : $g_i = MLP([\tau_\theta(c_i), s \cdot \gamma(p_i)] + (1 - s) \cdot e_i)$

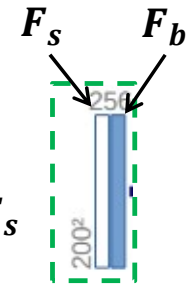
e_i is the learnable null token and and s is a binary value

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

ScaleU block



skip connections primarily contribute high-frequency features



main feature map F_b

lateral skip-connection features F_s

Theoretical basis

For instance-conditioned image generation, a notable improvement can be achieved by using channel-wise and learnable vectors to dynamically re-calibrate F_b and F_s

we introduce ScaleU, that has two *learnable*, *channel-wise* scaling vectors: $\mathbf{s}_b$, $\mathbf{s}_s$ for the main and skip-connected features, respectively. The main features $\mathbf{F}_b$ are scaled by a simple channel-wise multiplication: $\boldsymbol{F}'_b = \mathbf{F}_b \otimes (\tanh(\mathbf{s}_b) + 1)$. For the skip-connection features, we select the low-frequency (less than r_{thresh}) components using a frequency mask $\boldsymbol{\alpha}$ and scale them in the Fourier domain: $\mathbf{F}'_s = \text{IFFT}(\text{FFT}(\mathbf{F}_s) \odot \boldsymbol{\alpha})$. Here FFT$(\cdot)$ and IFFT$(\cdot)$ denote the Fast-Fourier and Inverse-Fast-Fourier transforms, $\odot$ is element-wise multiplication, and $\boldsymbol{\alpha}(r) = \tanh(\mathbf{s}_s) + 1$ if $r < r_{\text{thresh}}$ otherwise $= 1$, where r denotes the radius, and r_{thresh} refers to the threshold frequency. Both $\mathbf{s}_b$ and $\mathbf{s}_s$ are initially set to zero vectors.

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Multi-instance Sampler

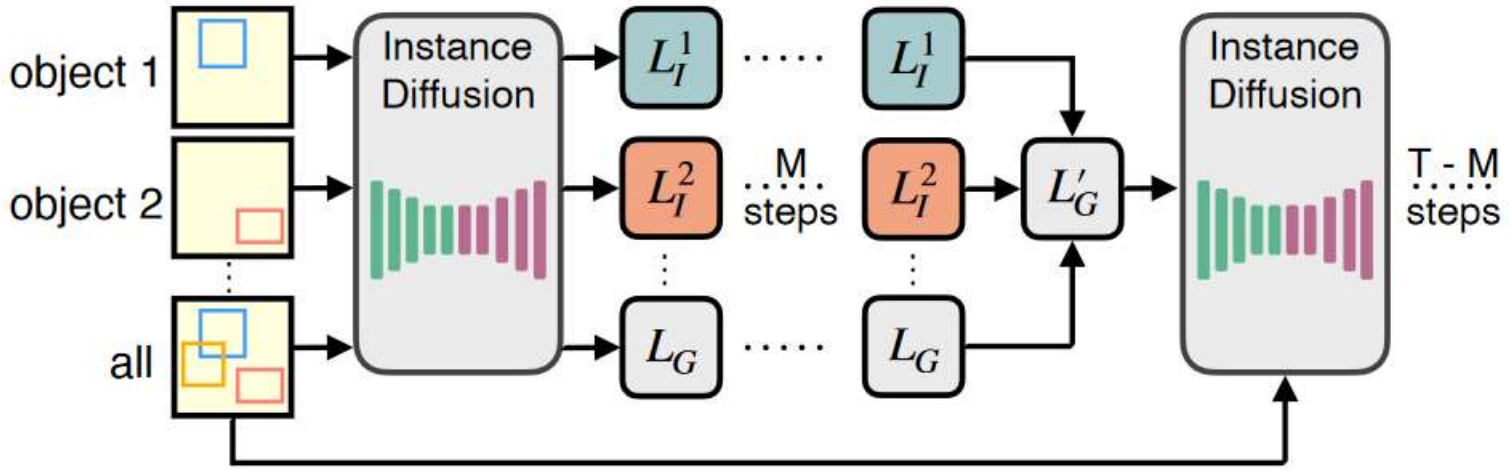


Figure 5. Model inference with **Multi-instance Sampler** to minimize information leakage across multiple instance conditionings.

Test data. We use standard benchmarks with bounding box and instance masks: 1) COCO [36] `val` with 80 classes; 2) large vocabulary instance segmentation dataset LVIS [19] `val` with over 1200 classes; 3) 250 selected samples (∼2 objects per image) from COCO `val` as in [28]. We do not use the real images from the dataset, and only use the text and location conditions. Notably, we also do not use any information from the `train` splits of the data which makes our evaluations zero-shot.

Bounding box. We follow prior work [25, 28, 34, 45] and use the YOLO score. Specifically, we use a pre-trained YOLOv8m-Det [25] detection model. We 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). We report AP_l^{box}, AP_m^{box}, and AP_s^{box}, which evaluate the model's performance based on different object sizes.

Instance mask. We compare YOLOv8m-Seg [25]'s detected instance masks in the generated image to the masks specified in the input using the COCO AP and AR metrics. To compare with [28], we report the IOU score for the mask.

Scribble. Since prior work has not reported on alignment performance using scribble, we introduced a new evaluation metric using YOLOv8m-Seg. We report "Points in Mask" (**PiM**), which measures how many of randomly sampled points in the input scribble lie within the detected mask.

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.

Compositional attribute binding. We 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".

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Location format input →	Boxes				Instance Masks					Points		Scribble	
Method	AP^{box}	AP^{box}_{50}	AR^{box}	FID (↓)	IoU	AP^{mask}	AP^{mask}_{50}	AR^{mask}	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	-	-	-	-	-	-	-	-	-
GLIGEN [34]*	19.3	34.6	31.1	-	-	-	-	-	-	-	-	30.2†	32.4†
ControlNet [65]‡	-	-	-	-	-	6.5	13.8	12.9	-	-	-	-	-
DenseDiffusion [28]	-	-	-	-	35.0 / 48.6	-	-	-	-	-	-	-	-
SpaText [4]‡	-	-	-	-	-	5.3	12.1	10.7	-	-	-	-	-
InstanceDiffusion	**38.8**	**55.4**	**52.9**	**23.9**	**61.6 / 71.4**	**27.1**	**50.0**	**38.1**	**25.5**	**81.1**	**27.5**	**72.4**	**27.3**
vs. prev. SoTA	**+19.2**	**+20.4**	**+21.8**	**-3.1**	**+25.4 / +22.8**	**+20.6**	**+36.2**	**+25.2**	-	-	-	**+42.2**	**-4.9**
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. *: evaluated with YOLOv8. †: 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). ‡: 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^{mask} using its official mask conditioned Image2Image generation pipeline. Hybrid: we add instance masks as additional conditions.

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Methods	Color		Texture		Human Eval
	$\text{Acc}^{\text{color}}$	$\text{CLIP}^{\text{local}}$	$\text{Acc}^{\text{texture}}$	$\text{CLIP}^{\text{local}}$	
GLIGEN	19.2	0.206	16.6	0.206	19.7
InstDiff	**54.4**	**0.250**	**26.8**	**0.225**	80.3
Δ	**+35.2**	**+0.044**	**+10.2**	**+0.019**	

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_{50}	AP_s	AP_m	AP_l	AP_r	AP_c	AP_f
Upper bound	44.6	57.7	33.2	55.0	66.1	31.4	44.5	50.5
GLIGEN [34]†	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
vs. prev. SoTA	**+8.0**	**+16.0**	**+3.9**	**13.7**	**+13.9**	**+5.3**	**+8.7**	**+8.4**

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. †: reproduced results.

point	box	mask	PiM	AP^{box}	$\text{AP}^{\text{box}}_{50}$	AP^{mask}	$\text{AP}^{\text{mask}}_{50}$
✓	✗	✗	81.1	-	-	-	-
✓	✓	✗	85.6	38.8	55.4	-	-
✓	✓	✓	**86.0**	**44.6**	**59.6**	**27.1**	**50.0**

Table 4. Multiple location formats at inference improves performance and helps the model to better respect location conditions.

#	*FA Fusion*	*MaskAttn*	*ScaleU*	*Inst. Cap.*	*MIS*	$\text{AP}^{\text{mask}}_{50}$	$\text{Acc}^{\text{color}}$	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**.

versions →	FreeU [51]	ScaleU
AP^{box}_{50}	52.2	55.4

(a) ScaleU

methods →	w/o extra tokens	w/ extra tokens
AP^{mask}_{50}	46.7	50.0

(b) extra tokens from binary masks

format →	polygons	+inside
AP^{mask}_{50}	47.5	50.0

(c) mask parameterization

# points →	64	128	256	512
AP^{mask}_{50}	45.7	48.5	50.0	50.0

(d) # points per mask

Table 6. Ablating design choices where the default settings are indicated in gray. **(a)** Compared to FreeU, our proposed ScaleU block improves the models ability to respect location conditions. **(b)** Using extra tokens from binary instances masks can improve the mask AP. **(c)** Parameterizing the instance masks using points on their boundaries and inside is beneficial. **(d)** Increasing the number of points used to parameterize masks improves performance.

Figure 6. Qualitative comparison of InstanceDiffusion *vs*. GLIGEN conditioned on multiple instance boxes and prompts. Prior work (bottom row) fails to accurately reflect specific instance attributes, *e.g.*, colors for the flower and puppies on the left, and not depicting a waterfall on the right. The generations also do not capture the correct instances, and are prone to information leakage across the instance prompts, *e.g.*, generating two similar instances on the right. InstanceDiffusion effectively mitigates these issues.

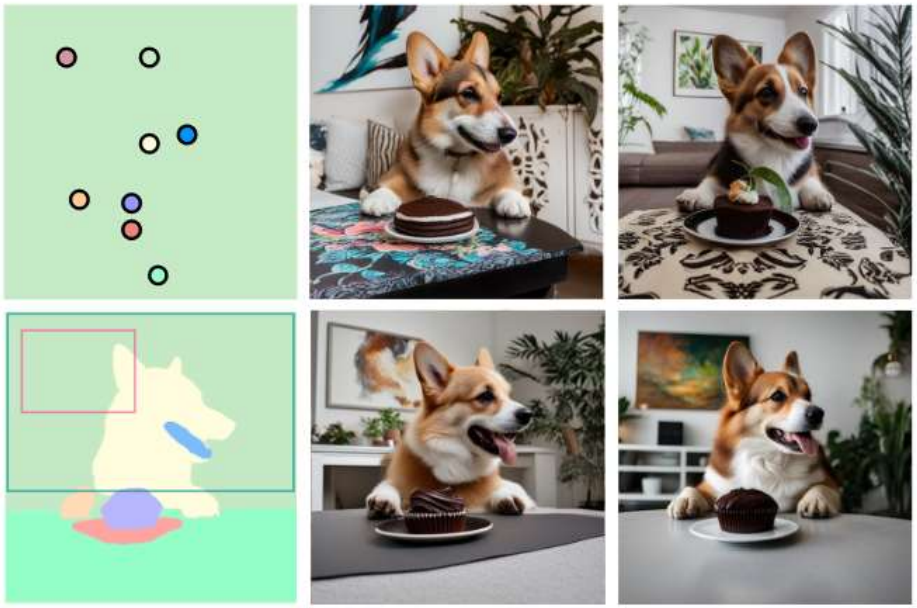


Image Caption: Cute Corgi at table in a living room with plants and painting on the wall. A chocolate cake is on the table. ***Instance Captions:*** 1) a Corgi sitting in front of a cupcake 2) Corgi's mouth and tongue 3) a plate 4) a chocolate cupcake on a plate 5) a white paw 6) a table 7) a living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

Figure 7. InstanceDiffusion image generation using various location conditions: points (row 1) and masks (row 2).

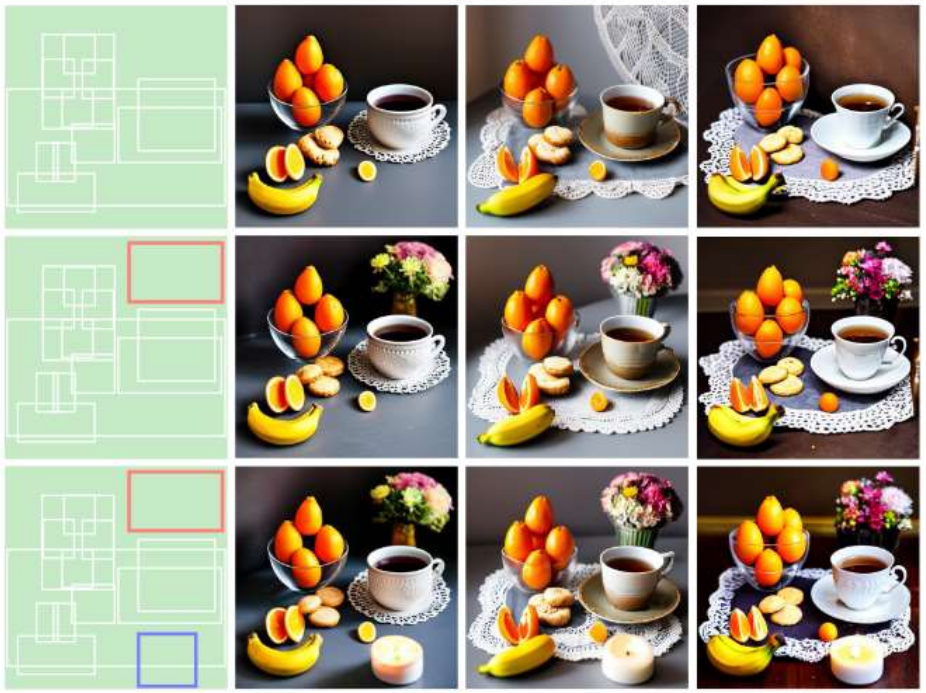


Image Caption: A cup of tea with tangerines, bananas, and cookies on the table. high quality. professional photo. ***Instance Captions:*** 1) a cup of tea on a lace doily 2) a close up of three oranges on a black background 3) oranges in a glass bowl on a table 4) a tray of pastries on a table with oranges 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

Figure 8. InstanceDiffusion can also support **iterative image generation**. Using the identical initial noise and image caption, InstanceDiffusion can progressively add new instances (like a bouquet of flowers in row two and a candle in row three), while minimally altering the pre-generated instances (row one). More results on iterative image generation that supports instance editing, replacing, moving and resizing can be found in appendix materials.