

# Grounding and Image Generation

GLIGEN  
ControlNet

CVPR 2023  
ICCV 2023

# Introduction



**Woo Chul Shin**

MSCS

**Interests:** Robotics,  
Dexterous Manipulation



**Mufei Li**

ML PhD

**Interests:** Memory  
Mechanisms of Foundation  
Models



**Alwin Jin**

MSCS

**Interests:**  
Post-training

# High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach\*, Andreas Blattmann\*, Dominik Lorenz, Patrick Esser, Bjorn Ommer

CVPR 2022

# Goal

- Give a quick recap of diffusion models and latent diffusion models
- Focus on how text conditioning works in LDM
- Set up the motivation by noting that GLIGEN and ControlNet build on this text conditioning mechanism

# Diffusion Model

- Forward process: gradually add Gaussian noise to data until it becomes nearly pure noise.
- Reverse process: train a neural network to iteratively denoise, step by step, recovering structure from noise.
- If we can learn the noise distribution at each step, we can sample new data by starting from noise and reversing the process.

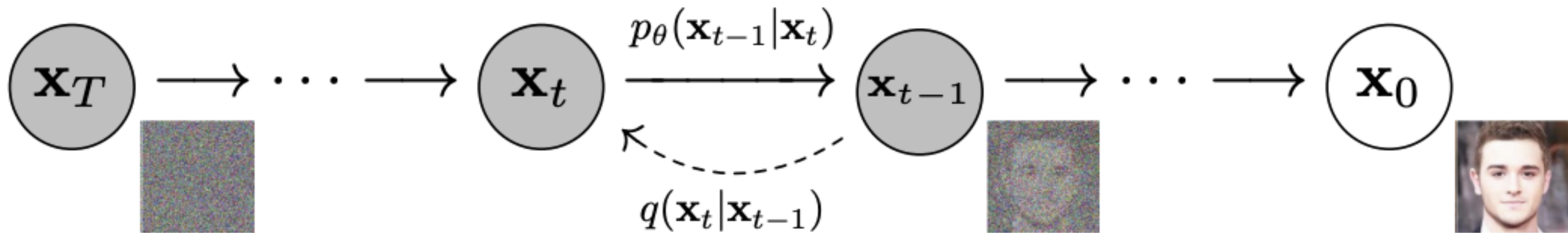
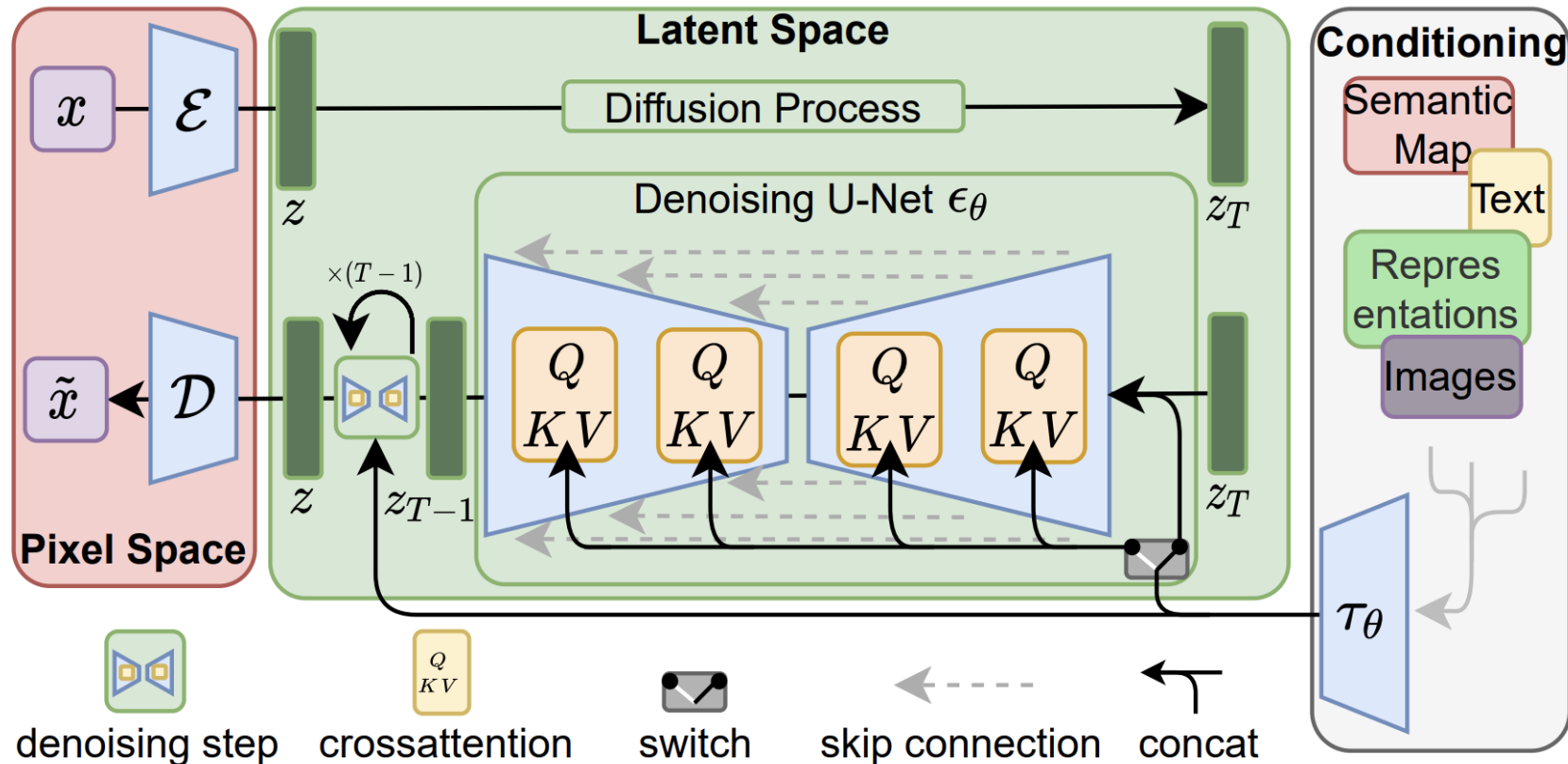


Figure 2: The directed graphical model considered in this work.

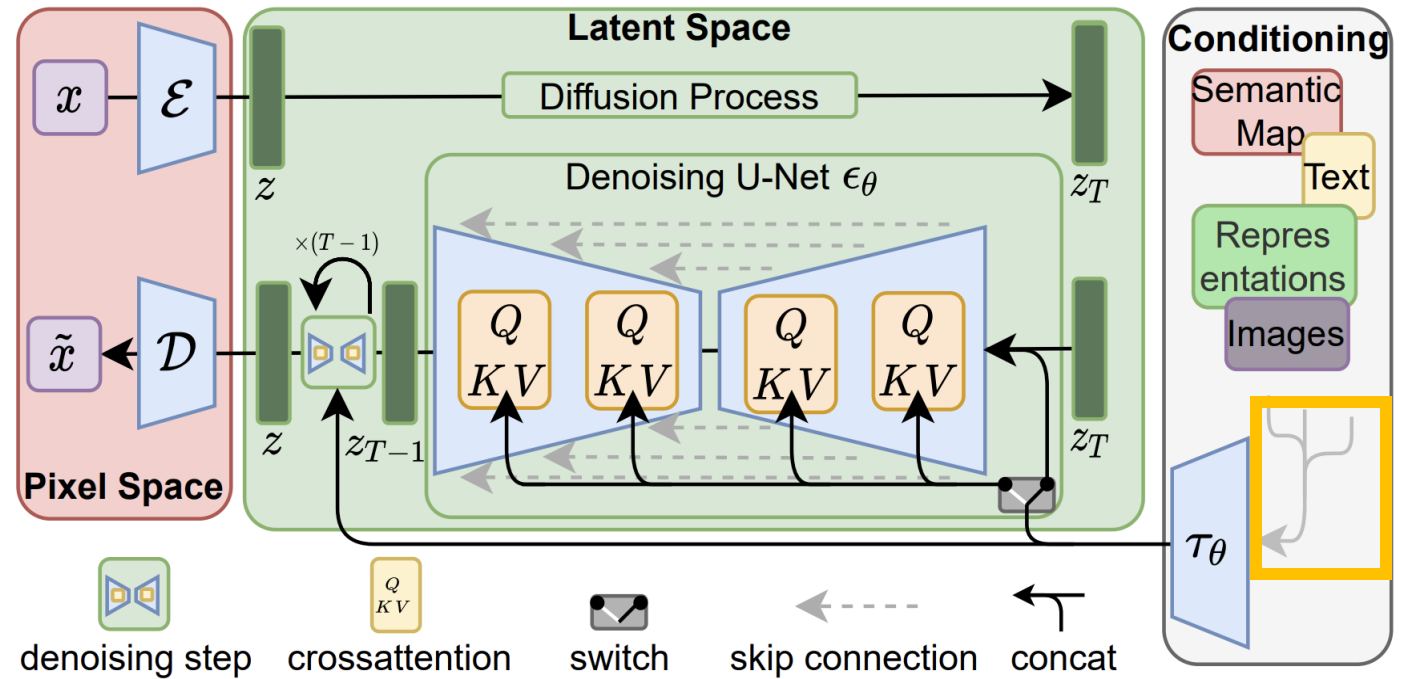
# Latent Diffusion Model

- Problem: pixel-space  $x_t$  is huge
- Two step approach
  1. Train encoder and decoder
  2. Diffusion in latent space
- Benefits
  - 8–16× smaller input size
  - Faster training and inference
  - U-Net models perceptual semantics, not raw pixels



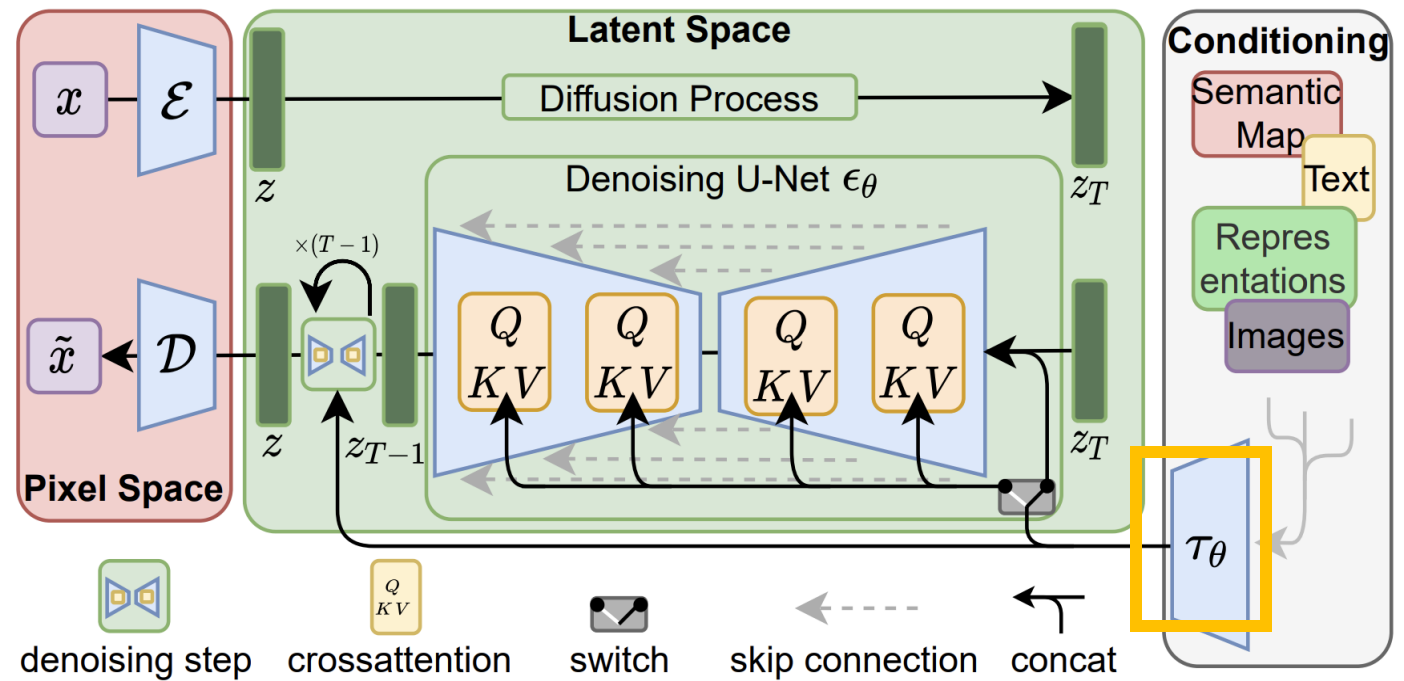
# Text conditioning in LDM

## 1. Tokenizer



# Text conditioning in LDM

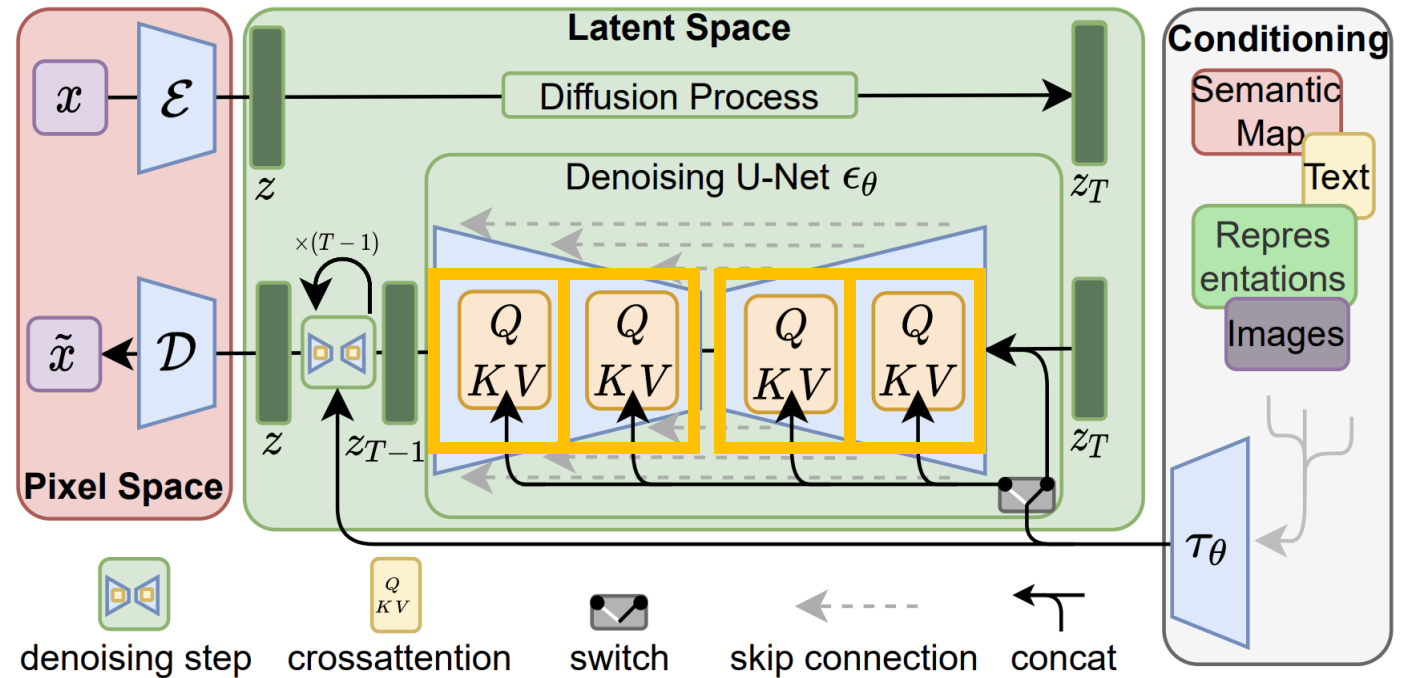
1. Tokenizer
2. Transformer





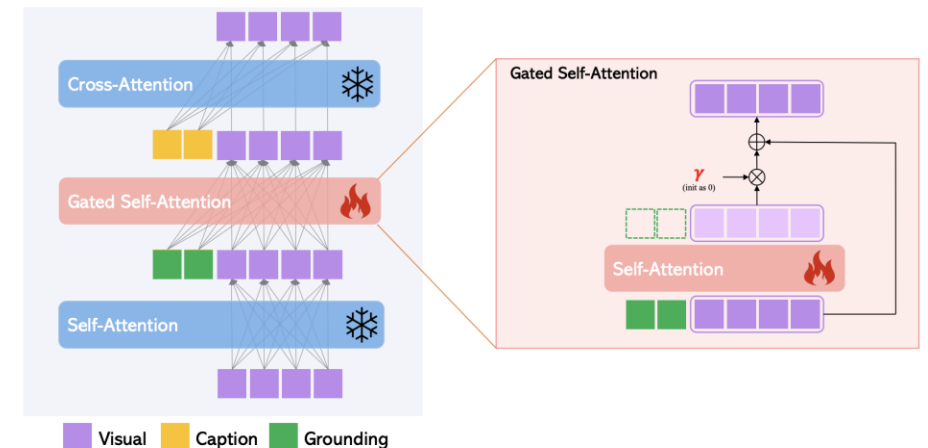
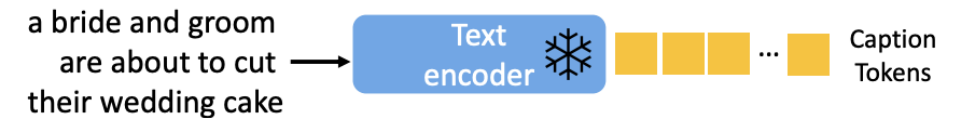
# Text conditioning in LDM

1. Tokenizer
2. Transformer
3. Cross-Attention with U-net's Intermediate layers
  - Q: flattened intermediate layer of U-net
  - K, V: encoded text prompt
  - Attention output is directly added back to the original input feature map



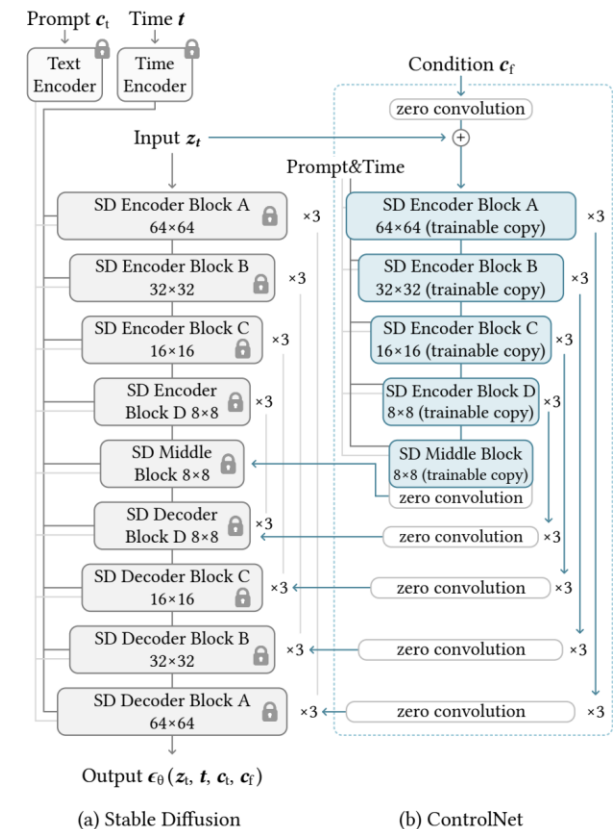
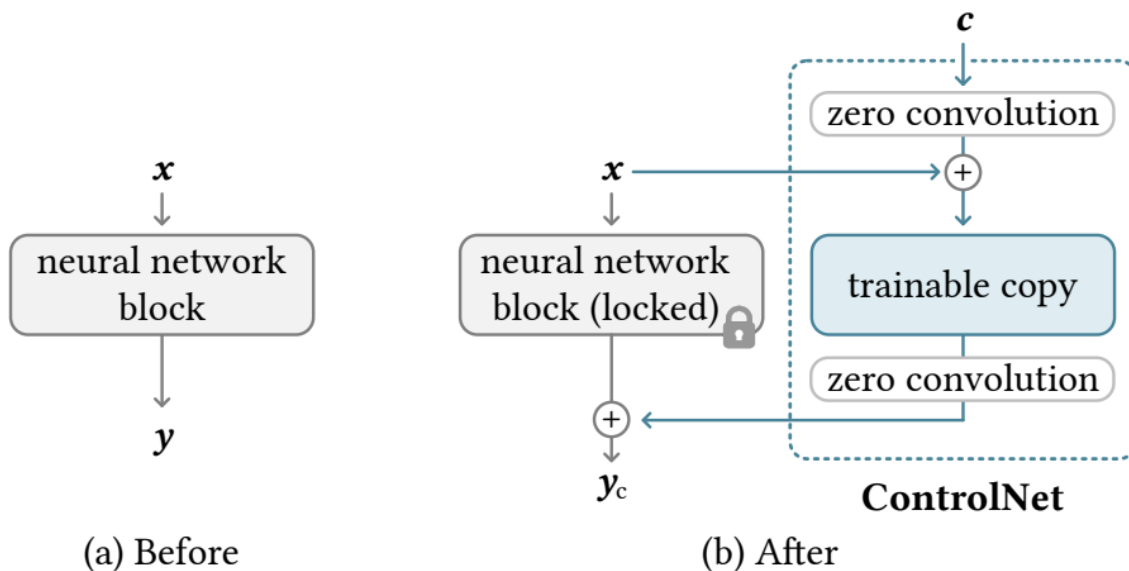
# Connection to GLIGEN

- Text conditioning in LDM has no explicit spatial grounding
  - e.g. A prompt like “a cat on the left and a dog on the right” may not respect spatial arrangement unless learned implicitly.
- GLIGEN introduces grounding tokens that tie text phrases to explicit regions of interest (Rols) in the image
  - Caption Tokens: CLIP embedding
  - Grounding Tokens:
    - Text token for the object
    - Bounding box -> MLP -> region embedding
- GLIGEN adds a learnable gate that decides how much influence the grounded tokens have compared to the plain caption tokens



# Connection to ControlNet

- ControlNet also builds on LDM text conditioning, but solves structural control (edges, depth, poses, etc.)
  - Base U-Net is frozen
  - A control branch (cloned U-Net) is added, initialized with zero-convs so it starts with no effect
  - Structural condition (e.g., Canny edges, pose maps) is passed into the control branch, which learns to output residual feature maps
  - Residuals are injected into the frozen base U-Net at multiple layers
  - Text conditioning is still done via cross-attention as in LDM



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

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, Yong Jae Lee

CVPR 2023



# Latent Diffusion Models Perform Text-to-Image Generation

## Text-to-Image Synthesis on LAION. 1.45B Model.

'A street sign that reads  
"Latent Diffusion" '

'A zombie in the  
style of Picasso'

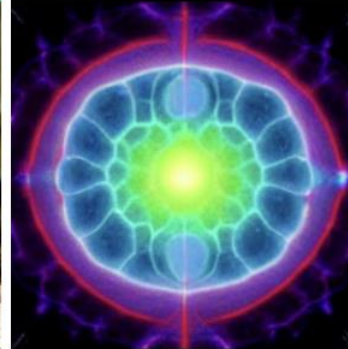
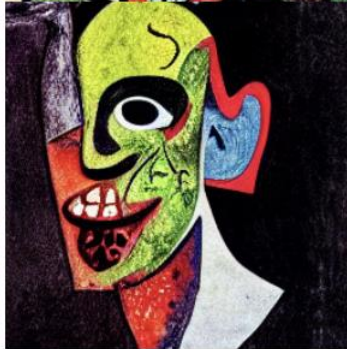
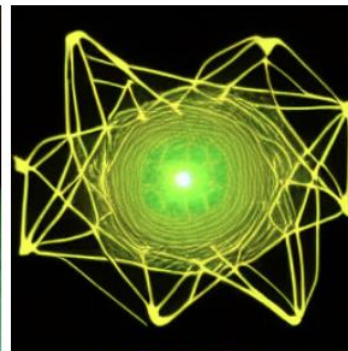
'An image of an animal  
half mouse half octopus'

'An illustration of a slightly  
conscious neural network'

'A painting of a  
squirrel eating a burger'

'A watercolor painting of a  
chair that looks like an octopus'

'A shirt with the inscription:  
"I love generative models!" '



# Motivation: Text Is Limited for Generation Conditioning

Text conditioning: A dog is on the left in the picture.

# Motivation: Text Is Limited for Generation Conditioning

Text conditioning: A dog is on the left in the picture.

*How large is the dog relative to the picture?*

*What is its precise coordinates?*

# Motivation: Text Is Limited for Generation Conditioning

Text conditioning: A dog is on the left in the picture.

ambiguous, imprecise, harming generation  
controllability

*How large is the dog relative to the picture?*

*What is its precise coordinates?*



# Motivation: Text Is Limited for Generation Conditioning

Text conditioning: A dog is on the left in the picture.

*How large is the dog relative to the picture?*

ambiguous, imprecise, harming generation  
controllability

*What is its precise coordinates?*

Visual information can be more expressive and precise!

# Motivation: Text Is Limited for Generation Conditioning

Text conditioning: A dog is on the left in the picture.

*How large is the dog relative to the picture?*

ambiguous, imprecise, harming generation  
controllability

*What is its precise coordinates?*

Visual information can be more expressive and precise!



(b)

Caption: "A dog / bird / helmet / backpack is on the grass"

Grounded image: red inset

# Motivation: Text Is Limited for Generation Conditioning

Text conditioning: A dog is on the left in the picture.

*How large is the dog relative to the picture?*

ambiguous, imprecise, harming generation  
controllability

*What is its precise coordinates?*

Visual information can be more expressive and precise!



(b)

Caption: "A dog / bird / helmet / backpack is on the grass"

Grounded image: red inset



1(d)

Caption: "a baby girl / monkey / Homer Simpson / is scratching her/its head"

Grounded keypoints: plotted dots on the left image

# Motivation: Text Is Limited for Generation Conditioning

Text conditioning: A dog is on the left in the picture.

*How large is the dog relative to the picture?*

ambiguous, imprecise, harming generation  
controllability

*What is its precise coordinates?*

Visual information can be more expressive and precise!



(b)

Caption: "A dog / bird / helmet / backpack is on the grass"

Grounded image: **red inset**



2(d)

Caption: "a baby girl / monkey / Homer Simpson / is scratching her/its head"

Grounded keypoints: **plotted dots on the left image**



(e)

Caption: "A vibrant colorful bird sitting on tree branch"

Grounded depth map: **the left image**

gia

# Most Related Work

DALL-E  
zero-shot  
text2image  
autoregressive

2021

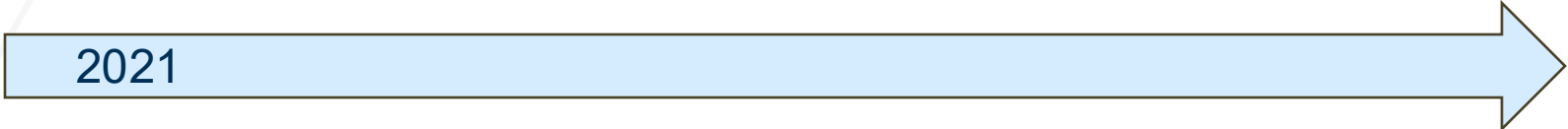


# Most Related Work

DALL-E

zero-shot  
text2image  
autoregressive

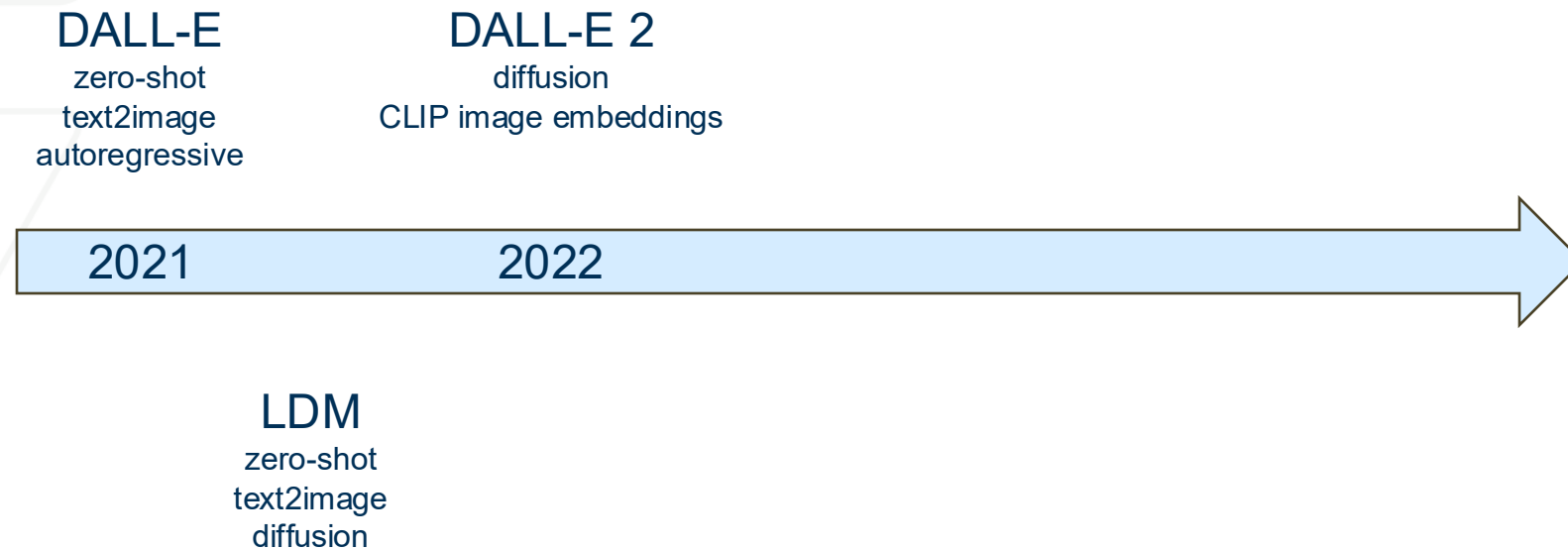
2021



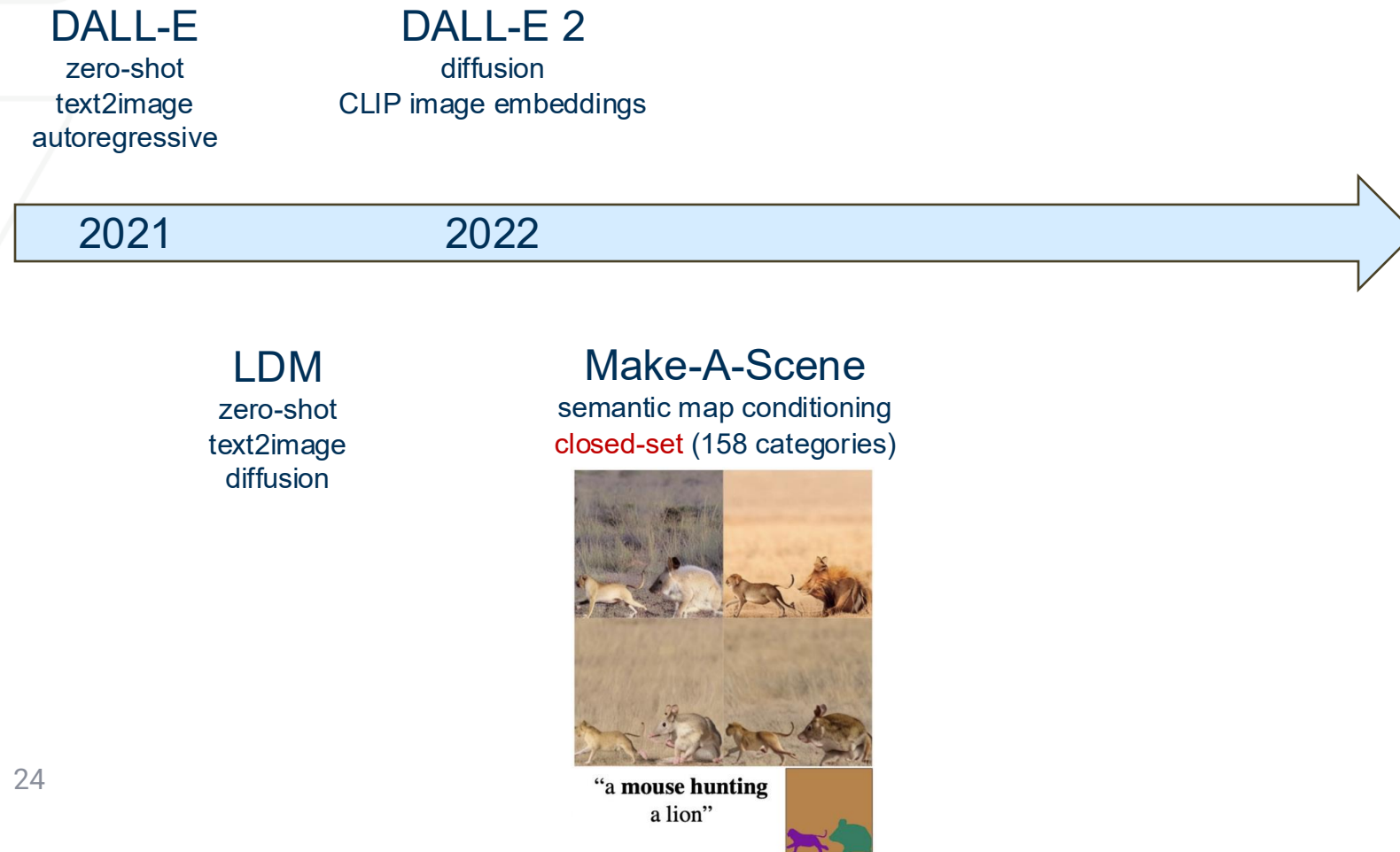
LDM

zero-shot  
text2image  
diffusion

# Most Related Work

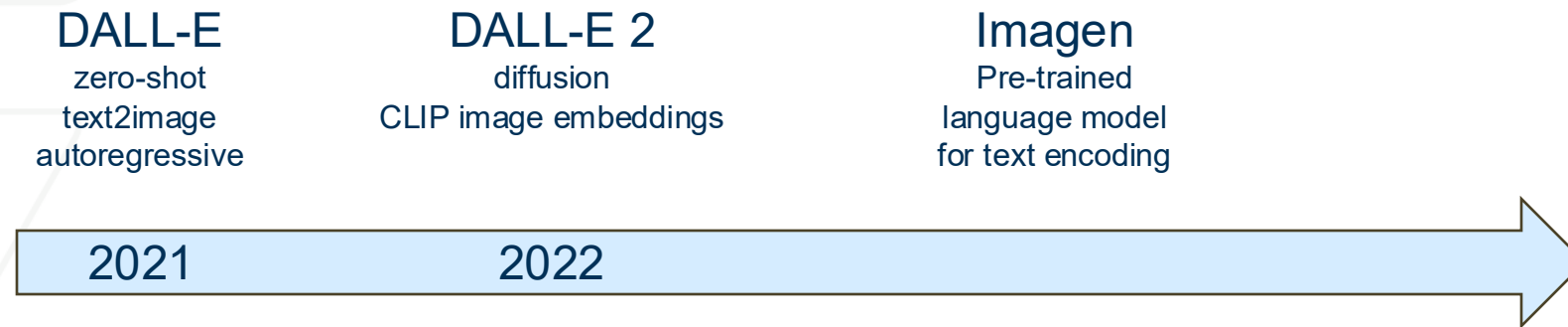


# Most Related Work



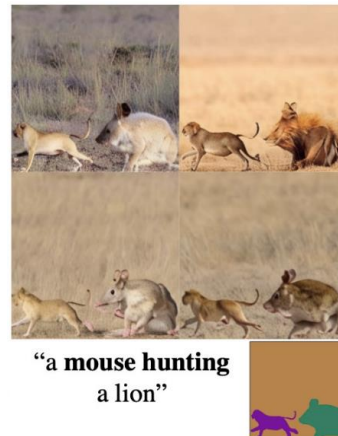


# Most Related Work

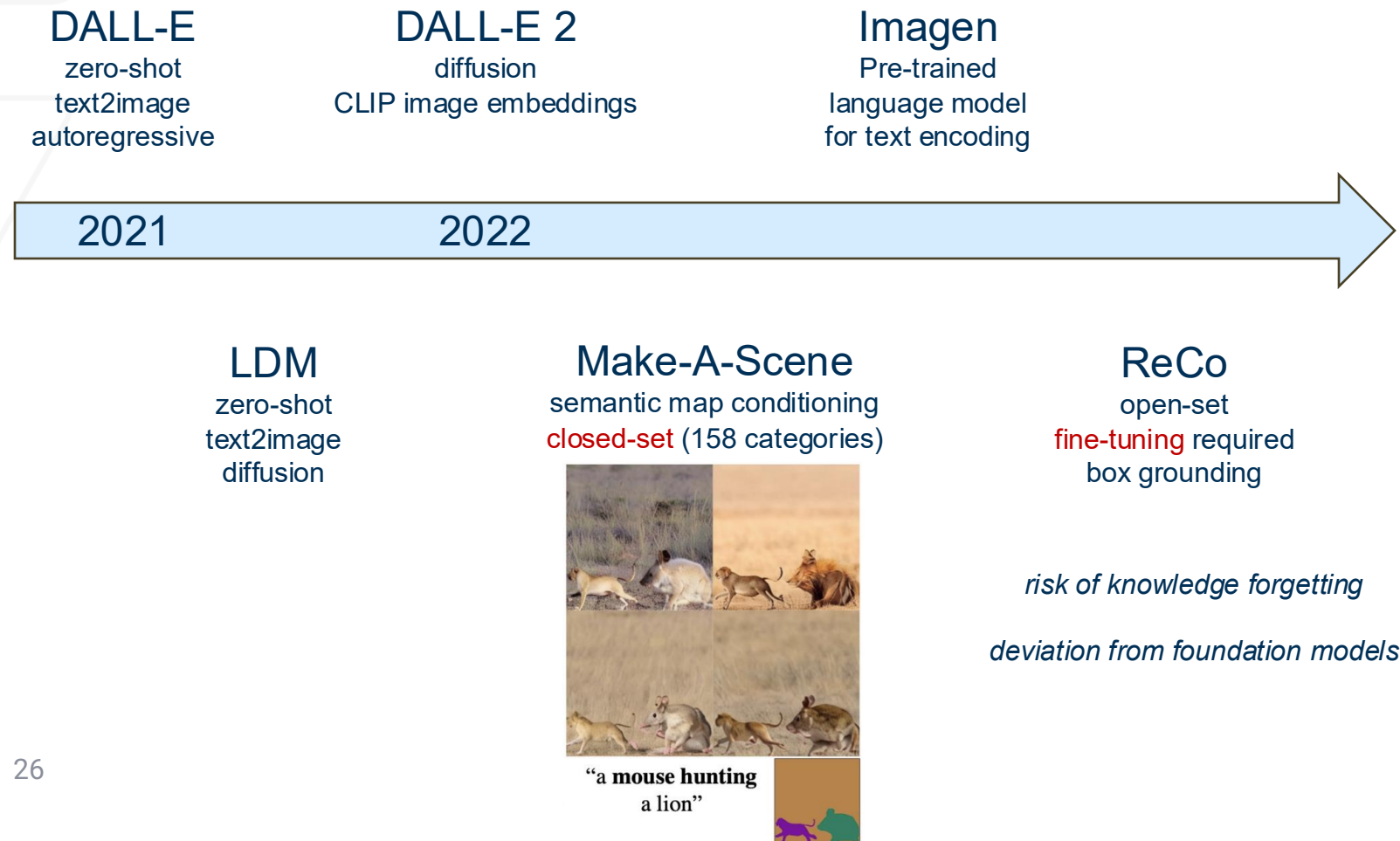


**LDM**  
zero-shot  
text2image  
diffusion

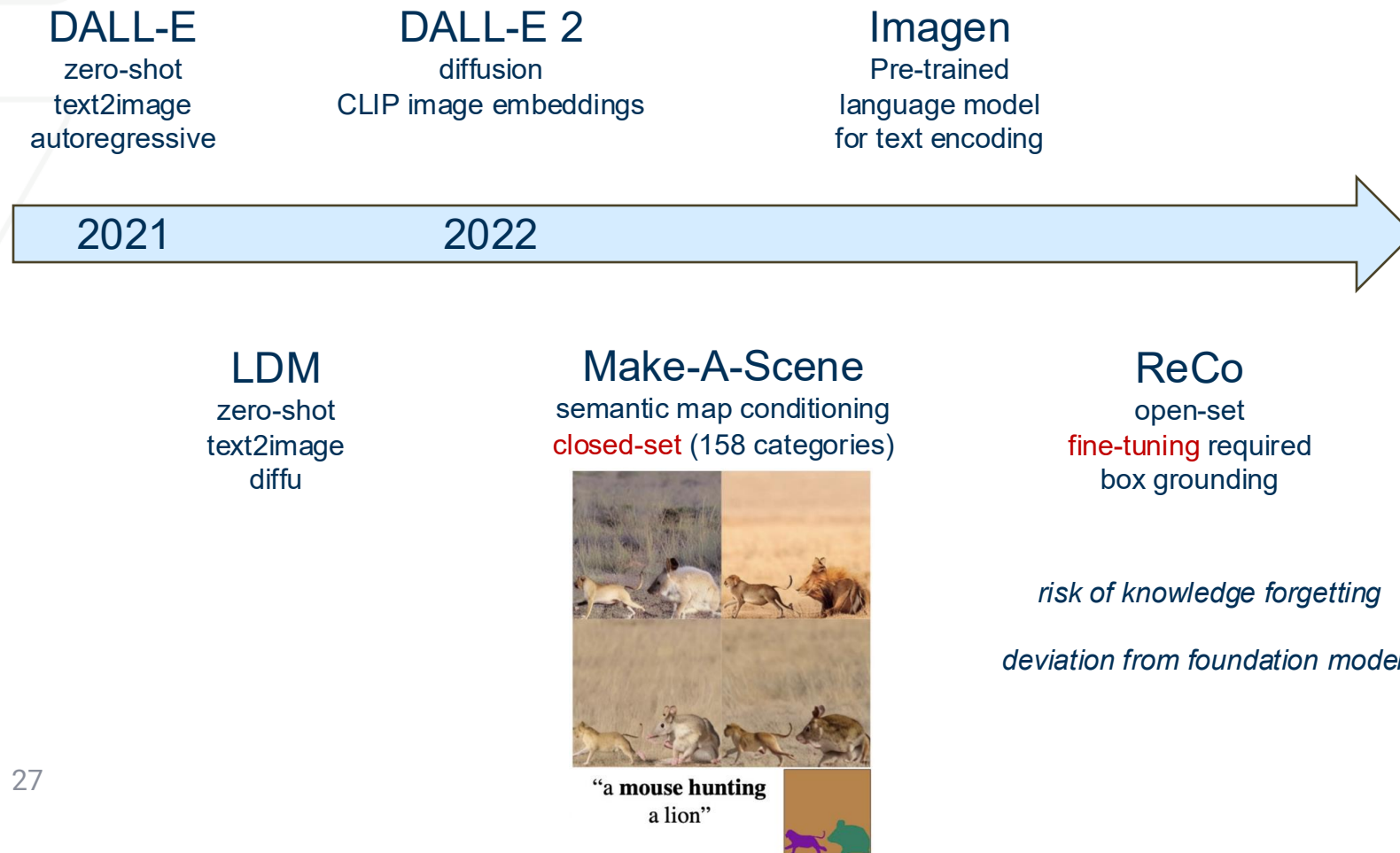
**Make-A-Scene**  
semantic map conditioning  
**closed-set** (158 categories)



# Most Related Work



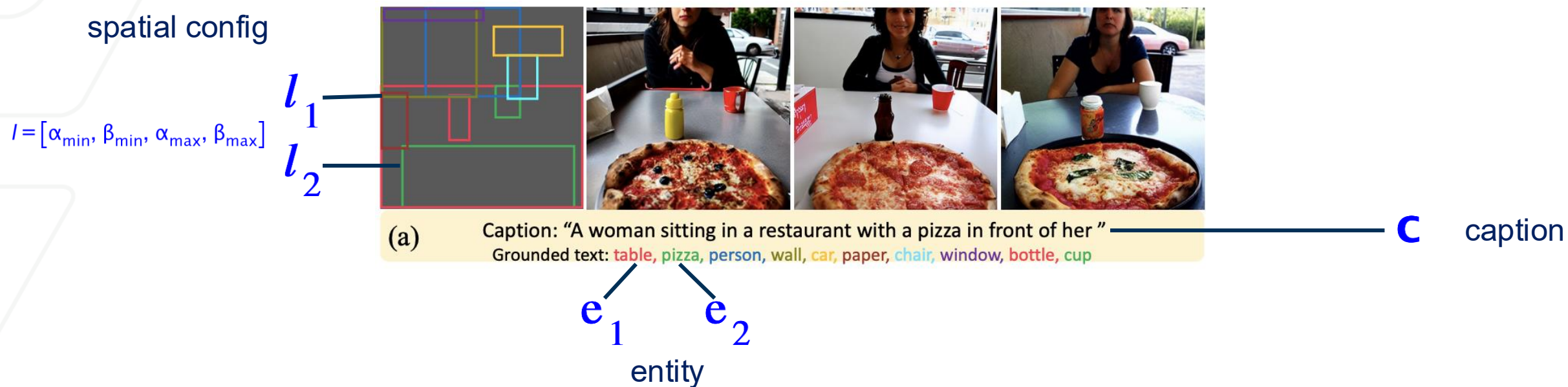
# Most Related Work



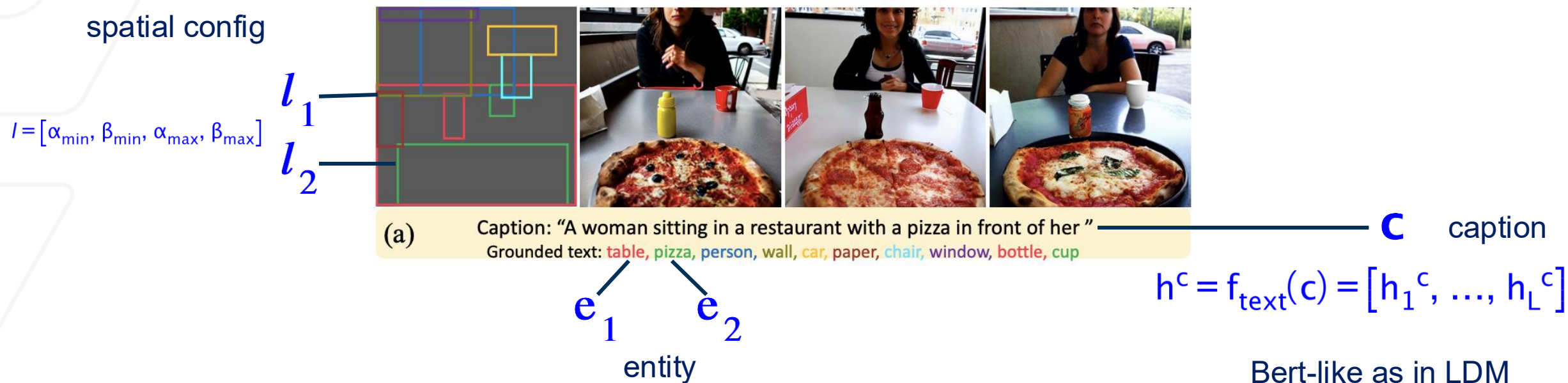
*Can we have:*

- open-set
- free of fine-tuning
- arbitrary visual conditioning

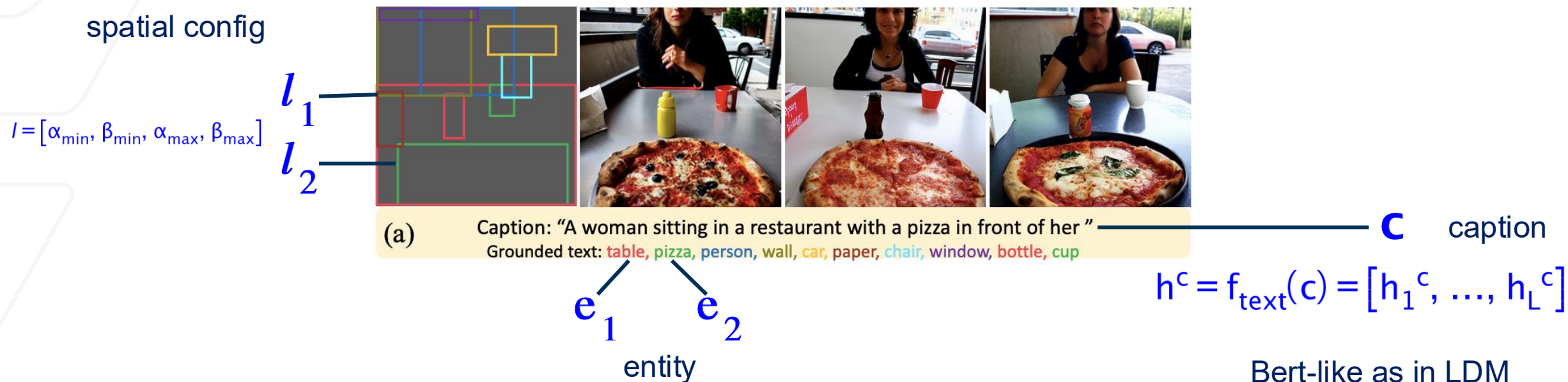
# Approach: Grounding Instruction Encoding



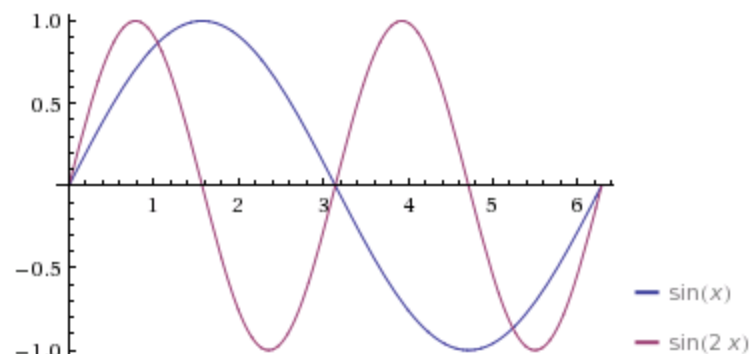
# Approach: Grounding Instruction Encoding



# Approach: Grounding Instruction Encoding

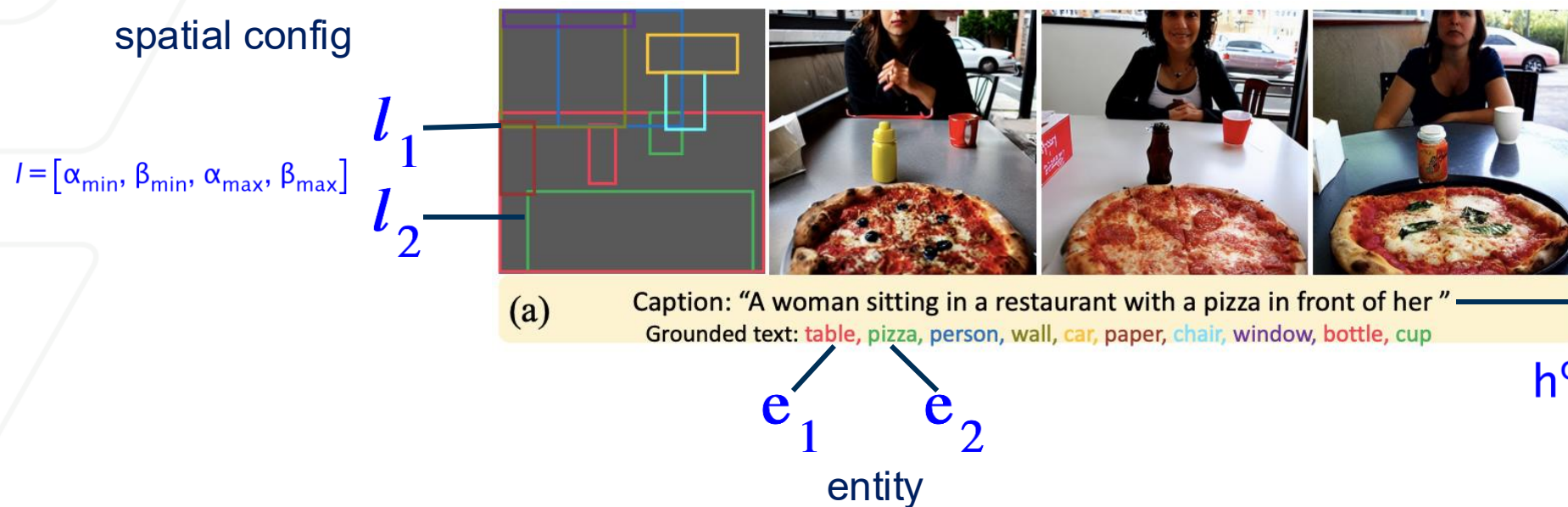


$$h^e = \text{MLP}(f_{\text{text}}(e), \text{Fourier}(l))$$





# Approach: Grounding Instruction Encoding



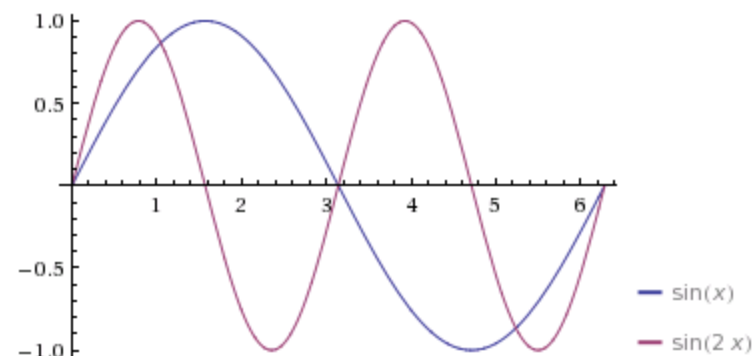
**C** caption

$$h^c = f_{\text{text}}(c) = [h_1^c, \dots, h_L^c]$$

Bert-like as in LDM

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

open-set  
compatibility

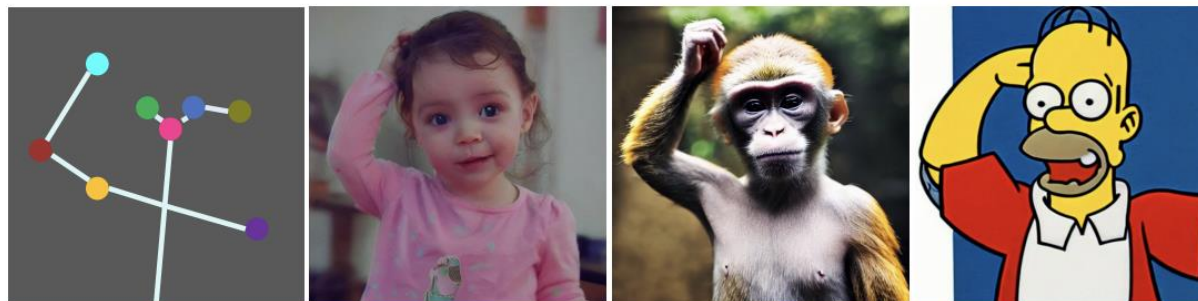


# Approach: Grounding Instruction Encoding

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

Compatible with other visual conditioning!

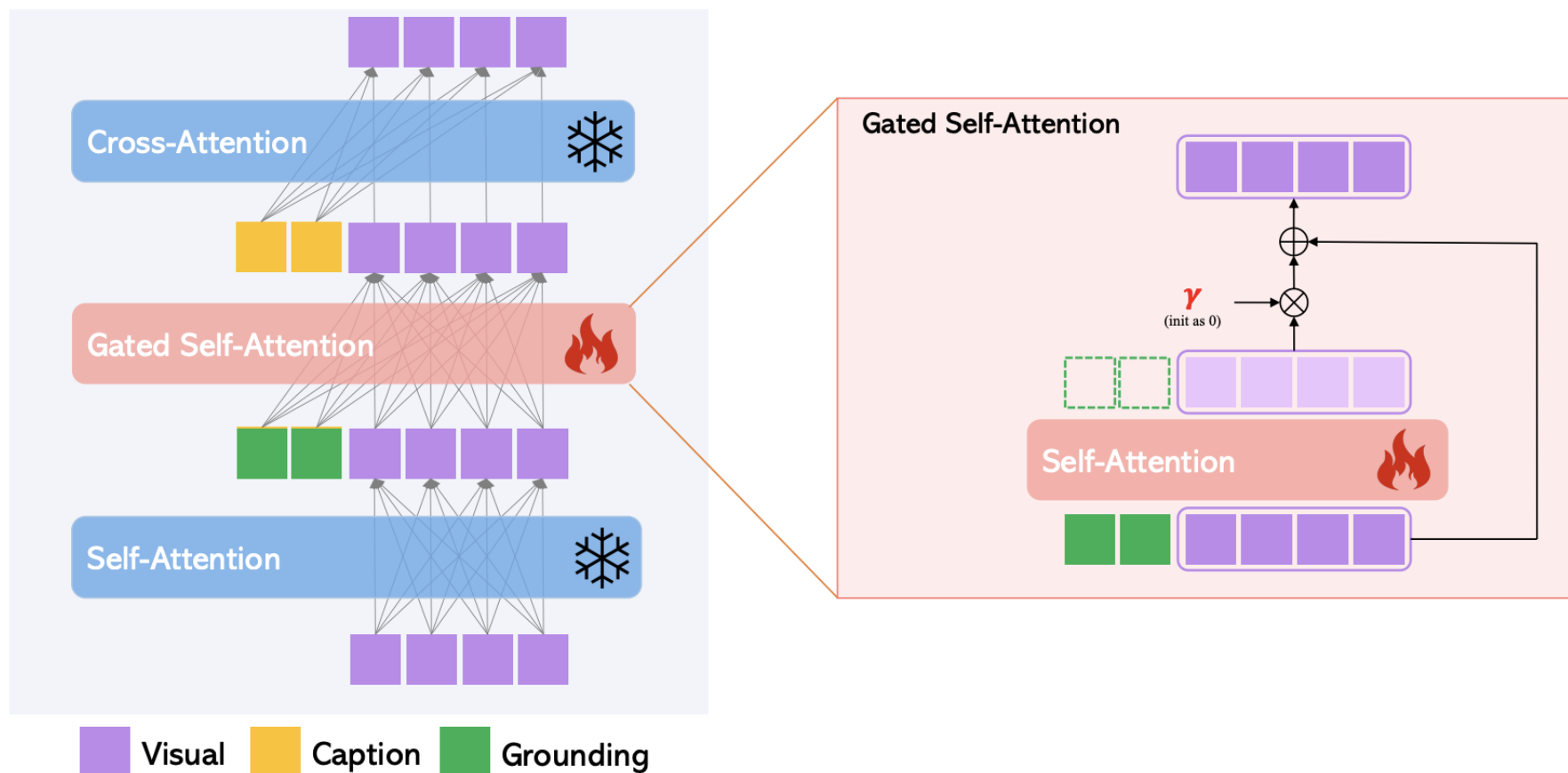
- Image prompt:  $f_{\text{image}}(e)$
- Keypoints:  $l = [x, y]$
- ...



(d) Caption: "a baby girl / monkey / Homer Simpson / is scratching her/its head"  
Grounded keypoints: **plotted dots on the left image**

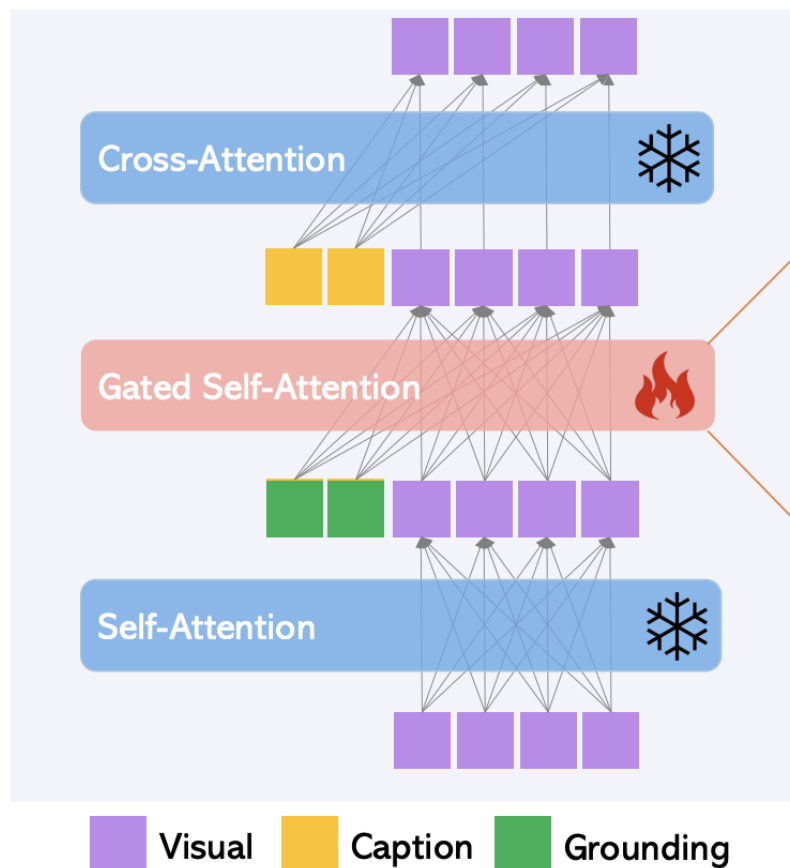


# Approach: Adaptation for Grounded Generation



The pre-trained model is fixed!

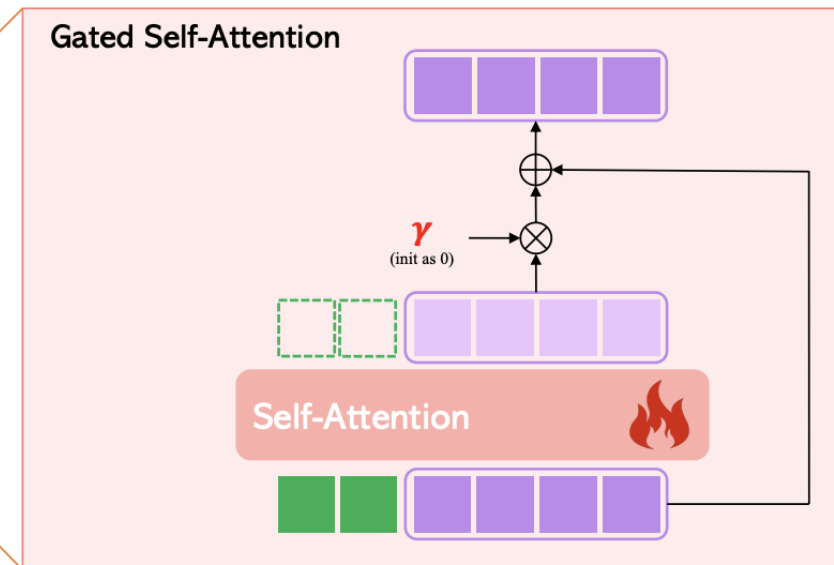
# Approach: Adaptation for Grounded Generation



$$\mathbf{v} = \mathbf{v} + \beta \cdot \tanh(\gamma) \cdot \text{TS}(\text{SelfAttn}([\mathbf{v}, \mathbf{h}^e]))$$

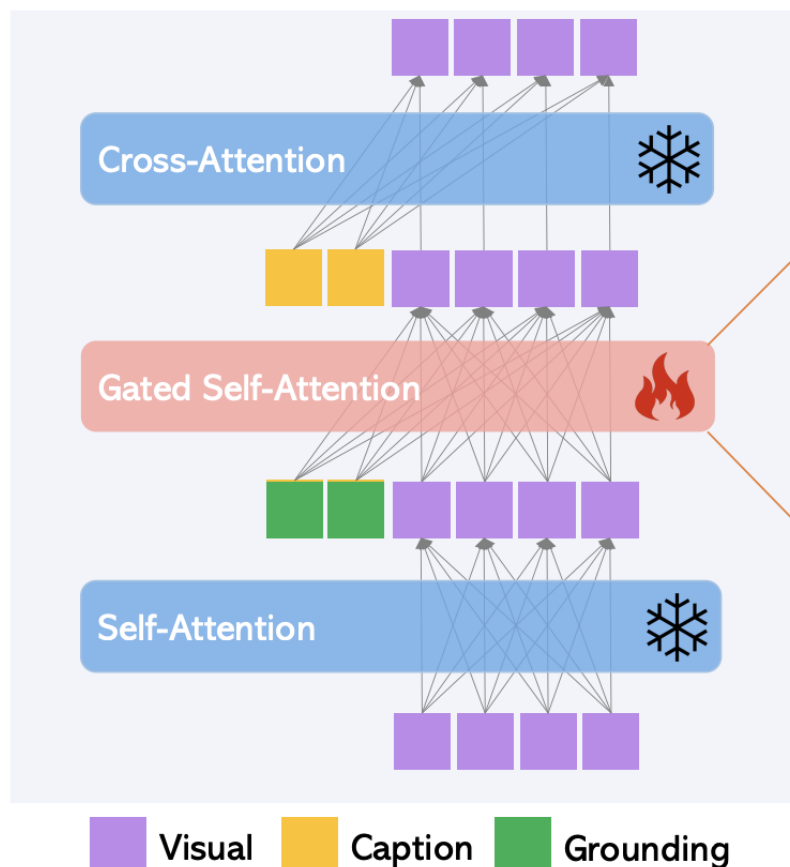
$\beta$  is 1 during training

$\gamma$  is learnable



The pre-trained model is fixed!

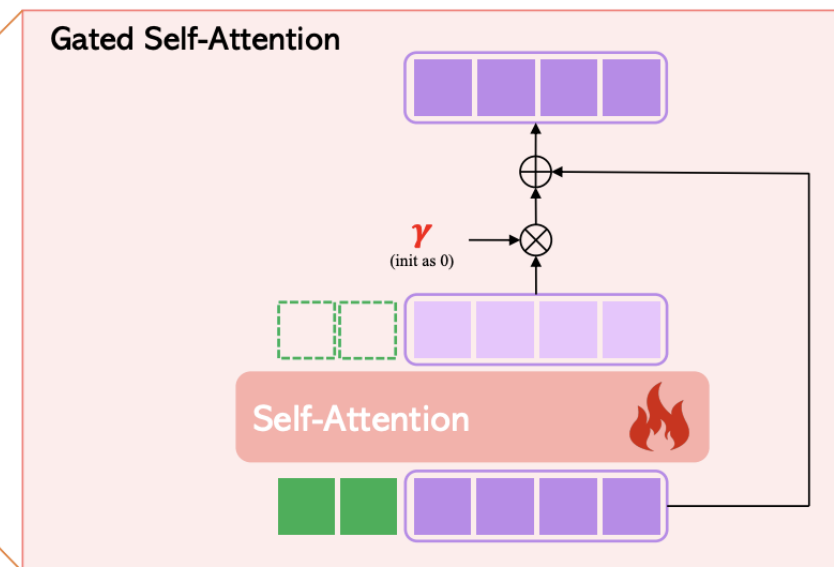
# Approach: Adaptation for Grounded Generation



$$\mathbf{v} = \mathbf{v} + \beta \cdot \tanh(\gamma) \cdot \text{TS}(\text{SelfAttn}([\mathbf{v}, \mathbf{h}^e]))$$

$\beta$  is 1 during training

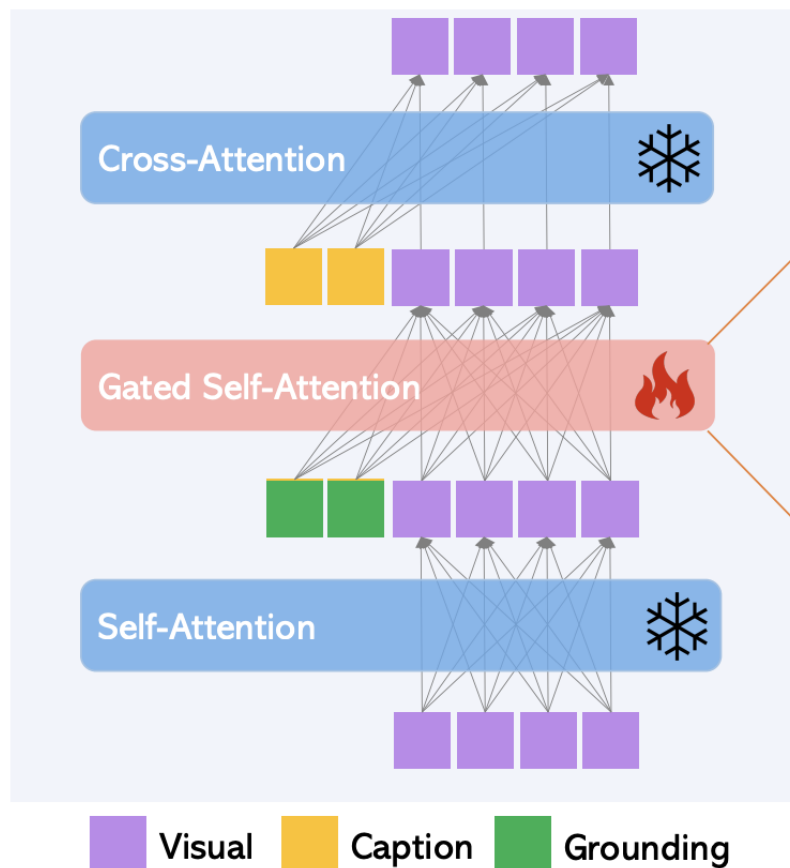
$\gamma$  is learnable



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

The pre-trained model is fixed!

# Approach: Adaptation for Grounded Generation

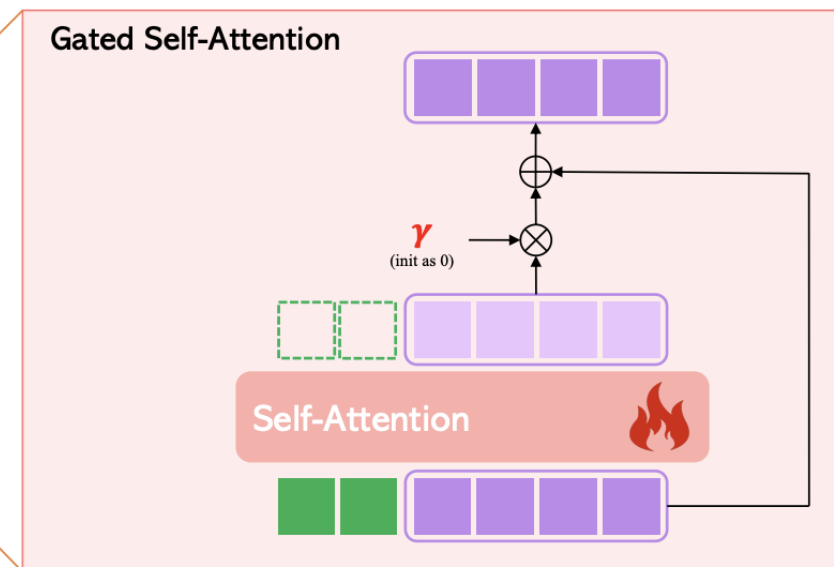


The pre-trained model is fixed!

$$\mathbf{v} = \mathbf{v} + \beta \cdot \tanh(\gamma) \cdot \text{TS}(\text{SelfAttn}([\mathbf{v}, \mathbf{h}^e]))$$

$\beta$  is 1 during training

$\gamma$  is learnable



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

Sampling  
schedule

$$\beta = \begin{cases} 1, & t \leq \tau * T \quad \# \text{ Grounded inference stage} \\ 0, & t > \tau * T \quad \# \text{ Standard inference stage} \end{cases}$$

# Experiment: Closed-Set Grounded Text2Img Generation

Model	Generation: FID (↓)		Grounding: YOLO (↑) AP/AP <sub>50</sub> /AP <sub>75</sub>
	Fine-tuned	Zero-shot	
CogView [11]	-	27.10	-
KNN-Diffusion [2]	-	16.66	-
DALL-E 2 [51]	-	10.39	-
Imagen [56]	-	7.27	-
Re-Imagen [7]	5.25	6.88	-
Parti [74]	3.20	7.23	-
LAFITE [82]	8.12	26.94	-
LAFITE2 [80]	4.28	8.42	-
Make-a-Scene [13]	7.55	11.84	-
NÜWA [69]	12.90	-	-
Frido [12]	11.24	-	-
XMC-GAN [77]	9.33	-	-
AttnGAN [70]	35.49	-	-
DF-GAN [65]	21.42	-	-
Obj-GAN [35]	20.75	-	-
LDM [53]	-	12.63	-
LDM*	5.91	11.73	0.6 / 2.0 / 0.3
GLIGEN (COCO2014CD)	5.82	-	21.7 / 39.0 / 21.7
GLIGEN (COCO2014D)	5.61	-	<b>24.0 / 42.2 / 24.1</b>
GLIGEN (COCO2014G)	6.38	-	11.2 / 21.2 / 10.7

Fréchet Inception Distance (FID):

1. Use pre-trained inception-v3 to embed images
2. Compare the two collections of real and generated images with a statistical distance

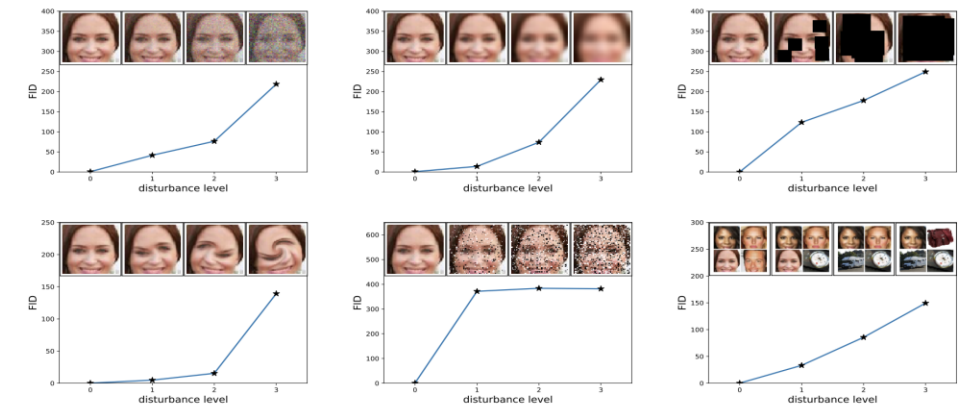


Figure 3: FID is evaluated for **upper left**: Gaussian noise, **upper middle**: Gaussian blur, **upper right**: implanted black rectangles, **lower left**: swirled images, **lower middle**: salt and pepper noise, and **lower right**: CelebA dataset contaminated by ImageNet images. The disturbance level rises from zero and increases to the highest level. The FID captures the disturbance level very well by monotonically increasing.

detection + caption data

detection data

pseudo box labels using GLIP for detection

# Experiment: Closed-Set Grounded Text2Img Generation

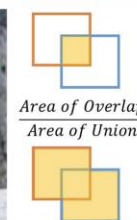
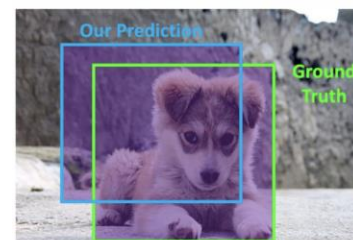
Model	Generation: FID (↓)		Grounding: YOLO (↑) AP/AP <sub>50</sub> /AP <sub>75</sub>
	Fine-tuned	Zero-shot	
CogView [11]	-	27.10	-
KNN-Diffusion [2]	-	16.66	-
DALL-E 2 [51]	-	10.39	-
Imagen [56]	-	7.27	-
Re-Imagen [7]	5.25	6.88	-
Parti [74]	3.20	7.23	-
LAFITE [82]	8.12	26.94	-
LAFITE2 [80]	4.28	8.42	-
Make-a-Scene [13]	7.55	11.84	-
NÜWA [69]	12.90	-	-
Frido [12]	11.24	-	-
XMC-GAN [77]	9.33	-	-
AttnGAN [70]	35.49	-	-
DF-GAN [65]	21.42	-	-
Obj-GAN [35]	20.75	-	-
LDM [53]	-	12.63	-
LDM*	5.91	11.73	0.6 / 2.0 / 0.3
GLIGEN (COCO2014CD)	5.82	-	21.7 / 39.0 / 21.7
GLIGEN (COCO2014D)	5.61	-	<b>24.0 / 42.2 / 24.1</b>
GLIGEN (COCO2014G)	6.38	-	11.2 / 21.2 / 10.7

Fréchet Inception Distance (FID):

1. Use pre-trained inception-v3 to embed images
2. Compare the two collections of real and generated images with a statistical distance

YOLO: Use a pre-trained YOLO-v4 to detect bounding boxes and compare them with the ground truth boxes using average precision.

IoU: Intersection over Union



Metrics	Metrics Meaning
AP	AP at IoU = 0.50: 0.05: 0.95
AP <sub>50</sub>	AP at IoU = 0.50
AP <sub>75</sub>	AP at IoU = 0.75

source:

[https://faculty.cc.gatech.edu/~zk15/teaching/AY2025\\_cs8803vlm\\_fall/L5\\_OpenVocabulary.pdf](https://faculty.cc.gatech.edu/~zk15/teaching/AY2025_cs8803vlm_fall/L5_OpenVocabulary.pdf)

detection + caption data

detection data

pseudo box labels using GLIP for detection

# Experiment: Closed-Set Grounded Text2Img Generation

Model	Generation: FID (↓)		Grounding: YOLO (↑) AP/AP <sub>50</sub> /AP <sub>75</sub>
	Fine-tuned	Zero-shot	
CogView [11]	-	27.10	-
KNN-Diffusion [2]	-	16.66	-
DALL-E 2 [51]	-	10.39	-
Imagen [56]	-	7.27	-
Re-Imagen [7]	5.25	6.88	-
Parti [74]	3.20	7.23	-
LAFITE [82]	8.12	26.94	-
LAFITE2 [80]	4.28	8.42	-
Make-a-Scene [13]	7.55	11.84	-
NÜWA [69]	12.90	-	-
Frido [12]	11.24	-	-
XMC-GAN [77]	9.33	-	-
AttnGAN [70]	35.49	-	-
DF-GAN [65]	21.42	-	-
Obj-GAN [35]	20.75	-	-
LDM [53]	-	12.63	-
LDM*	5.91	11.73	0.6 / 2.0 / 0.3
GLIGEN (COCO2014CD)	5.82	-	21.7 / 39.0 / 21.7
GLIGEN (COCO2014D)	5.61	-	<b>24.0 / 42.2 / 24.1</b>
GLIGEN (COCO2014G)	6.38	-	11.2 / 21.2 / 10.7

Fréchet Inception Distance (FID):

1. Use pre-trained inception-v3 to embed images
2. Compare the two collections of real and generated images with a statistical distance

YOLO: Use a pre-trained YOLO-v4 to detect bounding boxes and compare them with the ground truth boxes using average precision.

- Image synthesis quality is better than most SOTA baselines, and comparable to LDM<sup>^\*</sup>
- GLIGEN substantially outperforms LDM<sup>\*</sup> on grounding.
- COCO2014D has the overall best performance.

detection + caption data

detection data

pseudo box labels using GLIP for detection



# Experiment: Open-Set Grounded Text2Img Generation

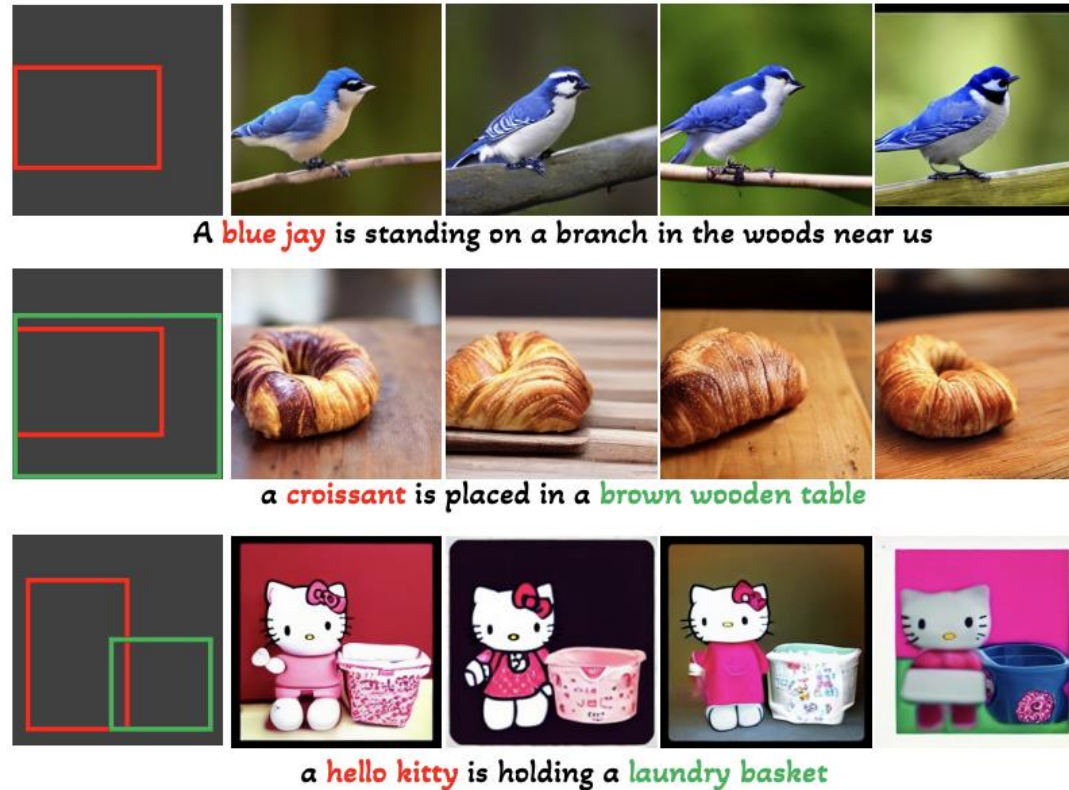


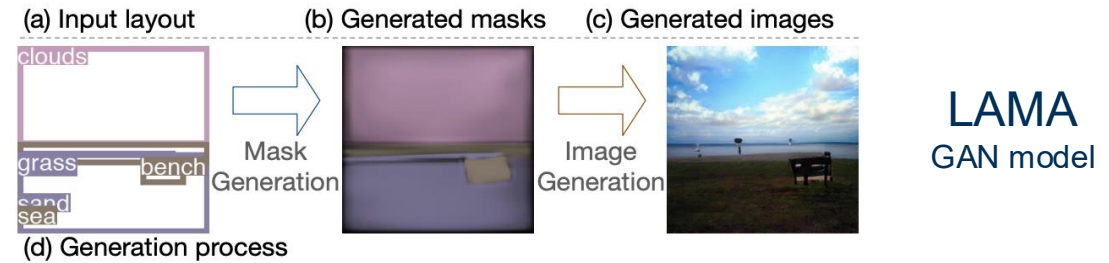
Figure 4. Our model can generalize to open-world concepts even when only trained using localization annotation from COCO.

benefit of pre-trained models



# Experiment: Open-Set Grounded Text2Img Generation

- $AP_r$ : Average precision for rare categories
- $AP_c$ : Average precision for common categories
- $AP_f$ : Average precision for frequent categories



Model	Training data	AP	$AP_r$	$AP_c$	$AP_f$
LAMA [40]	LVIS	2.0	0.9	1.3	3.2
GLIGEN-LDM	COCO2014CD	6.4	5.8	5.8	7.4
GLIGEN-LDM	COCO2014D	4.4	2.3	3.3	6.5
GLIGEN-LDM	COCO2014G	6.0	4.4	6.1	6.6
GLIGEN-LDM	GoldG,O365	10.6	5.8	9.6	13.8
GLIGEN-LDM	GoldG,O365,SBU,CC3M	11.1	9.0	9.8	13.4
GLIGEN-Stable	GoldG,O365,SBU,CC3M	10.8	8.8	9.9	12.6
Upper-bound	-	25.2	19.0	22.2	31.2

Outperforms LAMA  
(supervised baseline)  
on LVIS

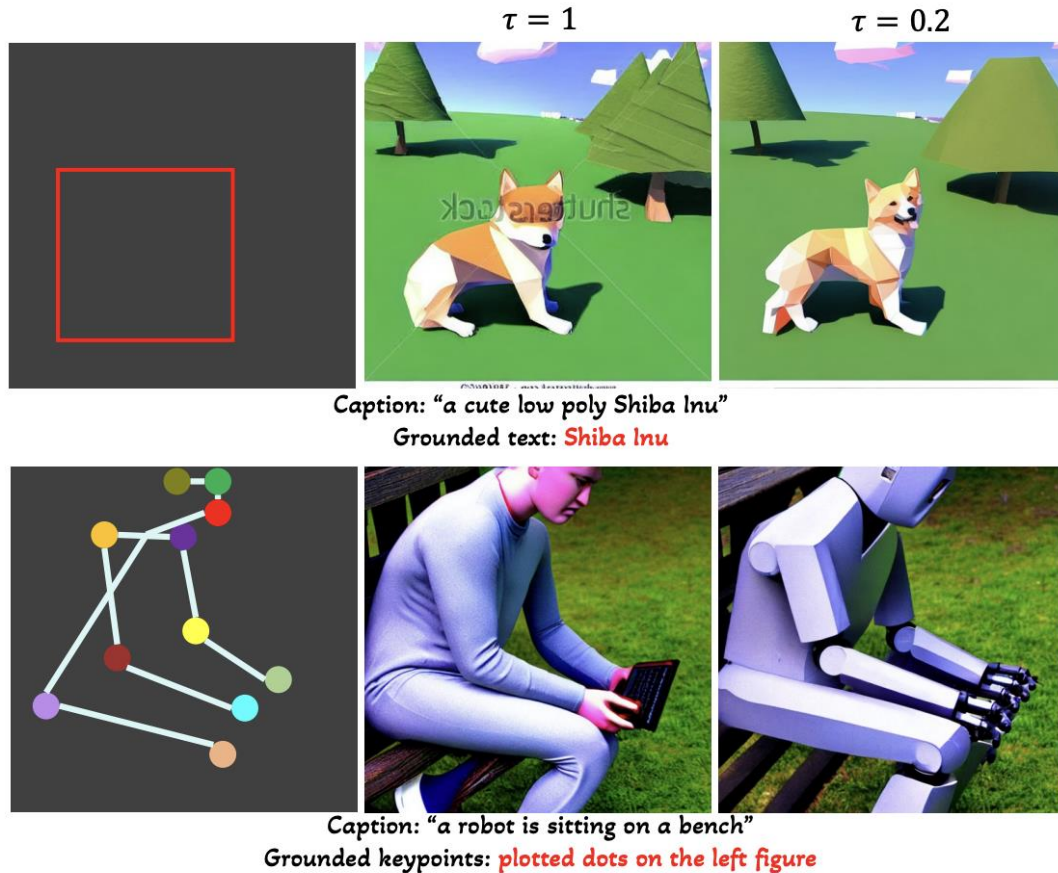
Scaling up the training  
data improves the  
performance.

# Experiment: Various Visual Conditioning



Figure 1. GLIGEN enables versatile grounding capabilities for a frozen text-to-image generation model, by feeding different grounding conditions. GLIGEN supports (a) text entity + box, (b) image entity + box, (c) image style and text + box, (d) keypoints, (e) depth map, (f) edge map, (g) normal map, and (h) semantic map.

# Experiment: Scheduled Sampling



$$\mathbf{v} = \mathbf{v} + \beta \cdot \tanh(\gamma) \cdot \text{TS}(\text{SelfAttn}([\mathbf{v}, \mathbf{h}^e]))$$

Sampling  
schedule

$$\beta = \begin{cases} 1, & t \leq \tau * T \quad \# \text{ Grounded inference stage} \\ 0, & t > \tau * T \quad \# \text{ Standard inference stage} \end{cases}$$

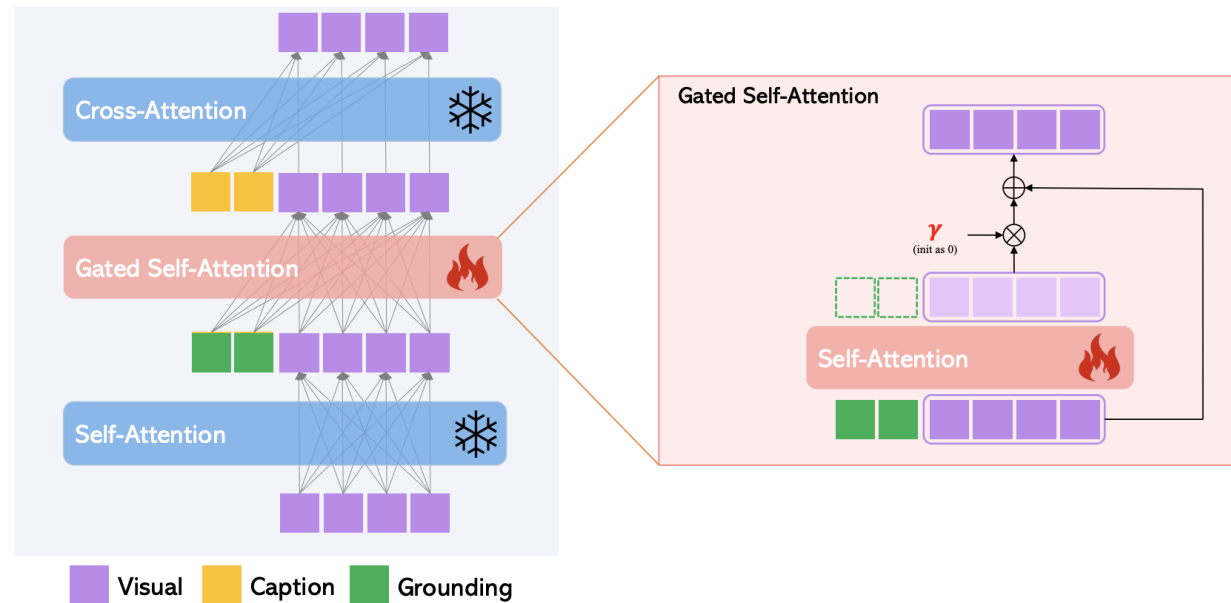


Figure 7. **Scheduled Sampling.** It can improve visual or extend a model trained in one domain (e.g., human) to the others.

# Reflection

## Strengths.

- ✓ First diffusion model compatible with various visual conditioning / grounding
- ✓ Open-Set
- ✓ Free of fine-tuning pre-trained models

# Reflection

## Strengths.

- ✓ First diffusion model compatible with various visual conditioning / grounding
- ✓ Open-Set
- ✓ Free of fine-tuning pre-trained models

## Limitations

- Entity-centric grounding rather than conceptual and contextual grounding
- Experiments primarily deal with bounding boxes
- Assumes a maximal input caption length and number of entities to ground



# Adding Conditional Control to Text-to-Image Diffusion Models

Lymin Zhang, Anyi Rao, Maneesh Agrawala

ICCV 2023

# Brief Recap: Text-to-Image Diffusion

## Text-to-Image Synthesis on LAION. 1.45B Model.

'A street sign that reads  
"Latent Diffusion" '

'A zombie in the  
style of Picasso'

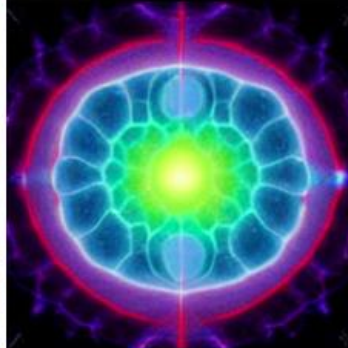
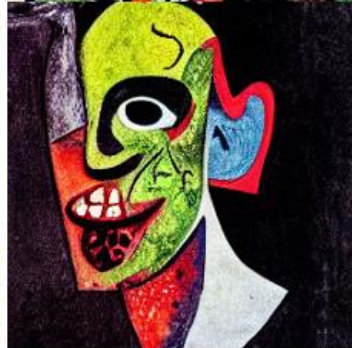
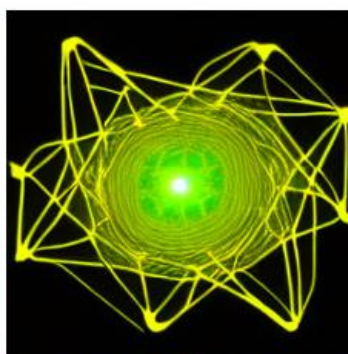
'An image of an animal  
half mouse half octopus'

'An illustration of a slightly  
conscious neural network'

'A painting of a  
squirrel eating a burger'

'A watercolor painting of a  
chair that looks like an octopus'

'A shirt with the inscription:  
"I love generative models!" '





# Brief Recap: Text-to-Image Diffusion

## Text-to-Image Synthesis on LAION. 1.45B Model.

'A street sign that reads  
"Latent Diffusion" '

'A zombie in the  
style of Picasso'

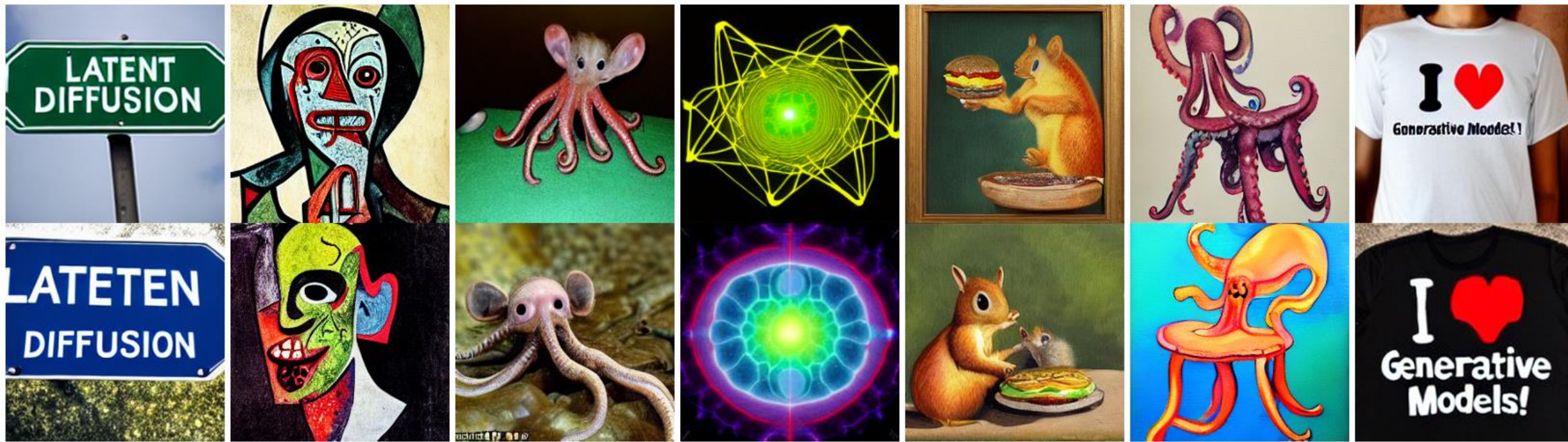
'An image of an animal  
half mouse half octopus'

'An illustration of a slightly  
conscious neural network'

'A painting of a  
squirrel eating a burger'

'A watercolor painting of a  
chair that looks like an octopus'

'A shirt with the inscription:  
"I love generative models!" '



Only conditioned on text!



# Brief Recap: GLIGEN



# Brief Recap: GLIGEN



Still text-conditioned! (+ bounding boxes)

# Motivation: Image-Based Spatial Conditioning

- Detailing exact spatial compositions is hard with only text
- Grounding enables high level composition only
- Consistency challenges



# Motivation: Image-Based Spatial Conditioning

- Detailing exact spatial compositions is hard with only text
- Grounding enables high level composition only
- Consistency challenges

What if we could condition on images too?



Input Canny edge



Default



"masterpiece of fairy tale, giant deer, golden antlers"



"..., quaint city Galic"

# ControlNet

**Idea: fine-tune existing model for image-based spatial conditioning**

**Q: why might this not work?**

- Catastrophic forgetting
- Mode collapse

Dog example results



Subject images

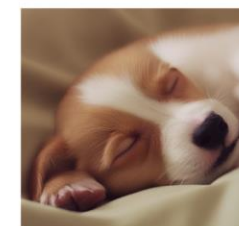
Fine-tuning



...on the beach



...in a bucket



...sleeping soundly

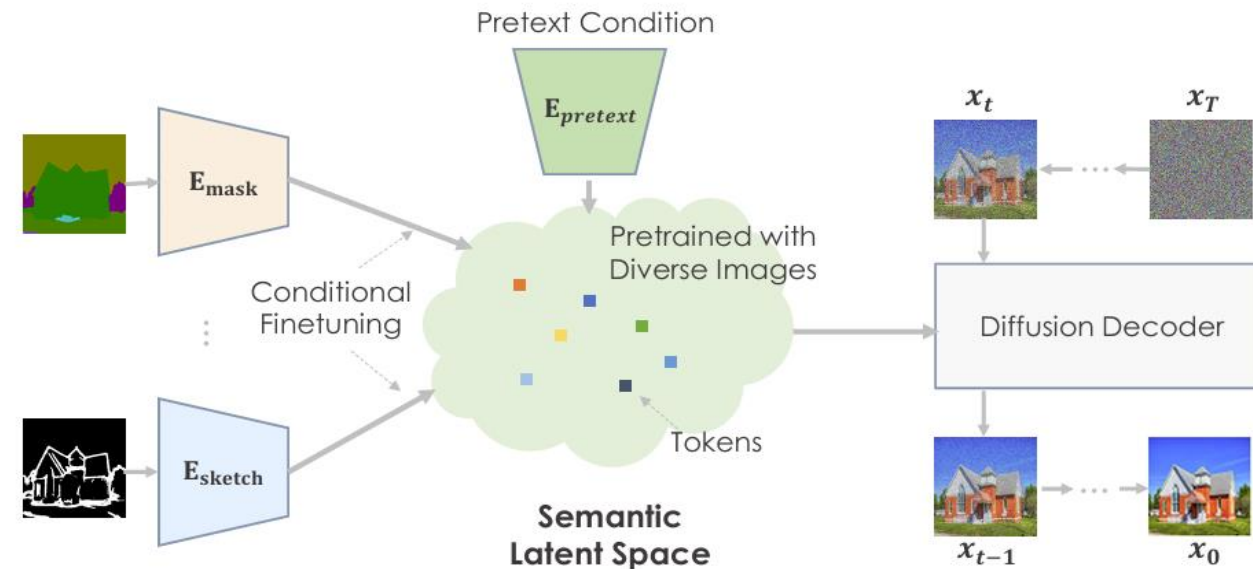


...as a superhero

# Related Work: Image-to-Image Translation

## Pretraining is All You Need (PITI)

- Historically I2I is done with GANs
- Use large pretrained diffusion model
- Fine-tune task-specific adapters for downstream tasks

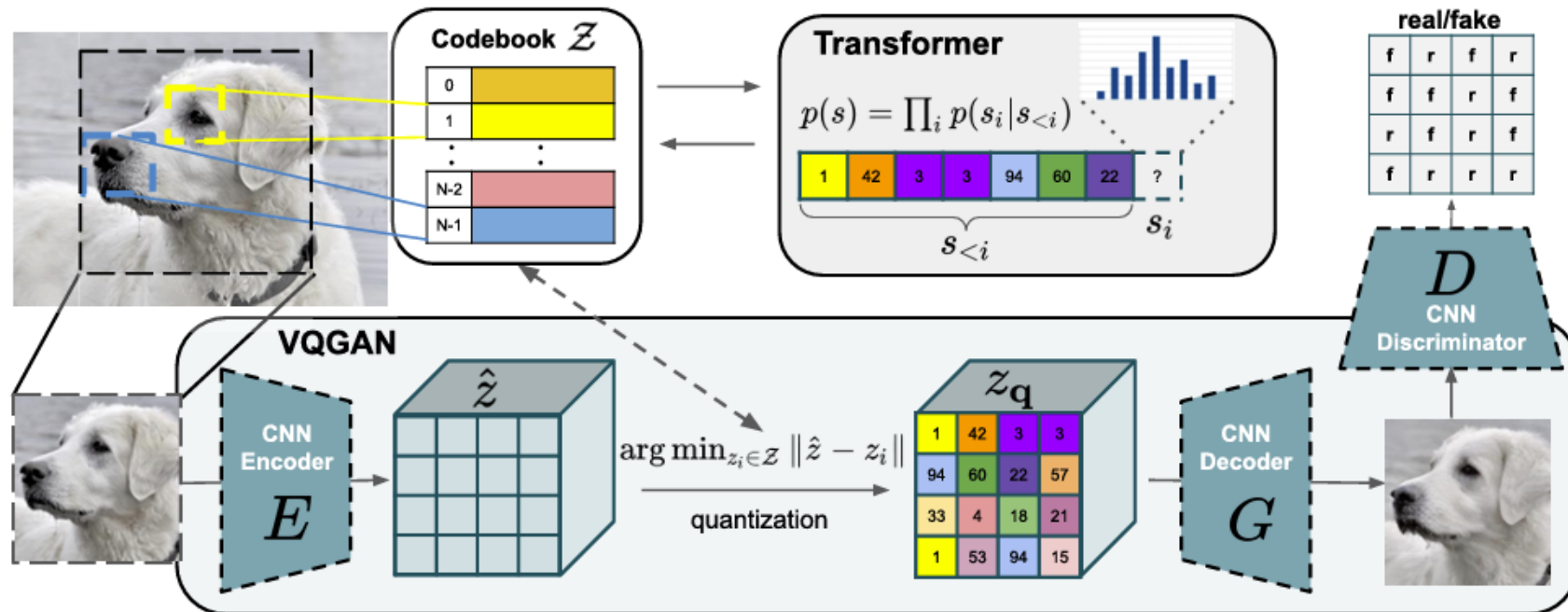




# Related Work: Image-to-Image Translation

## Taming Transformers for Image Synthesis

- Vision transformer I2I approach
- Use a convolutional VQGAN to learn a discrete codebook
- Use transformer to model code sequences
- Reconstruct code sequences back to image

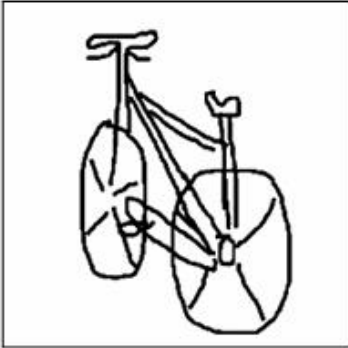


# Related Work: Image-to-Image Translation

## Sketch-Guided Diffusion

- Given sketch and text prompt, guide image generation with the sketch
- Learn an auxiliary network that predicts sketch images
- During denoising, use this network to guide image generation
- Only supports sketch guidance

Input Sketch



"A photo of a bicycle"



"An origami bicycle"



"A bicycle in a snowy weather"



"A macro photo of a toy bicycle"



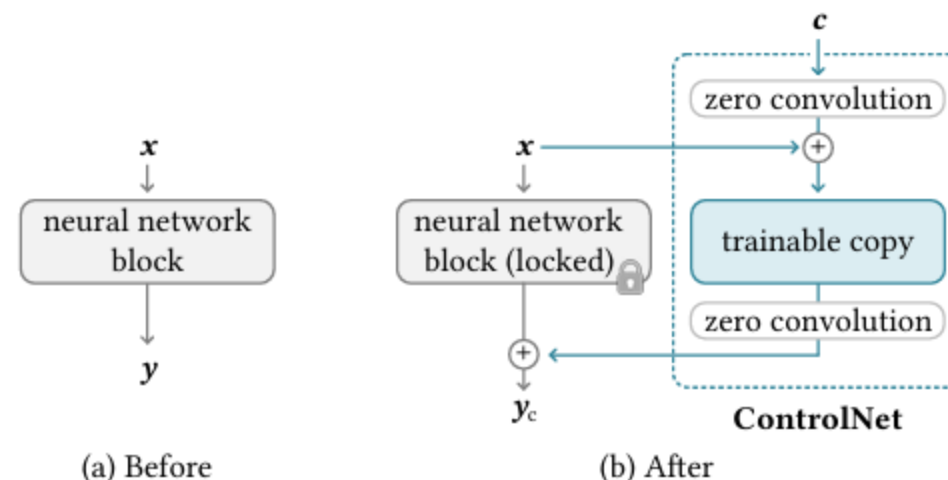
"A bicycle made of wood"



# ControlNet

## Freeze core model and add a "conditioning branch"

- Freeze the original NN block and make a trainable copy
- Add zero convolution layers (weights are zero)
- Zero convolution layer weights eventually become non-zero

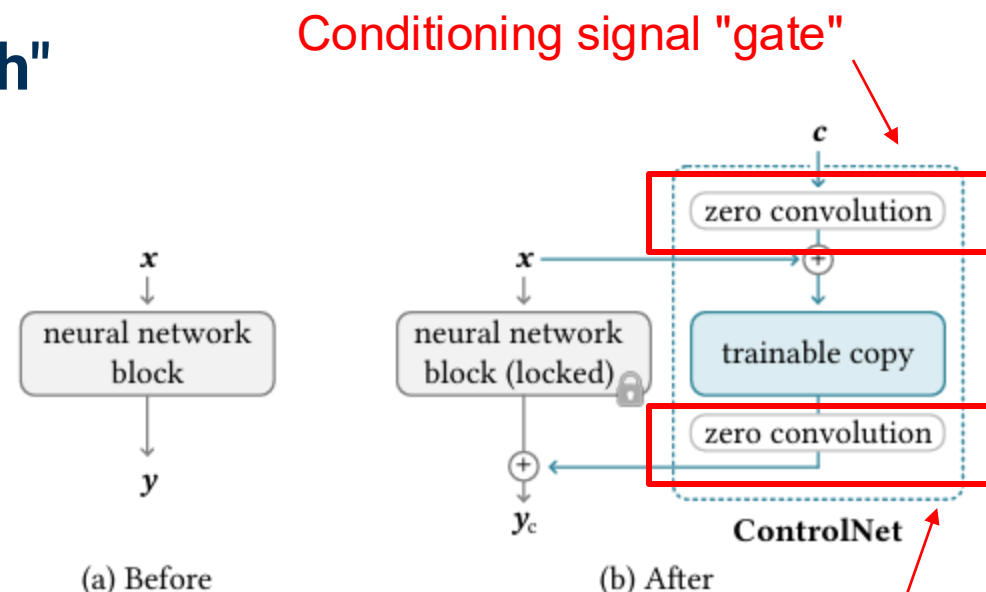


Over time, the conditioning branch learns how much of the conditioning signal to inject!

# ControlNet

## Freeze core model and add a "conditioning branch"

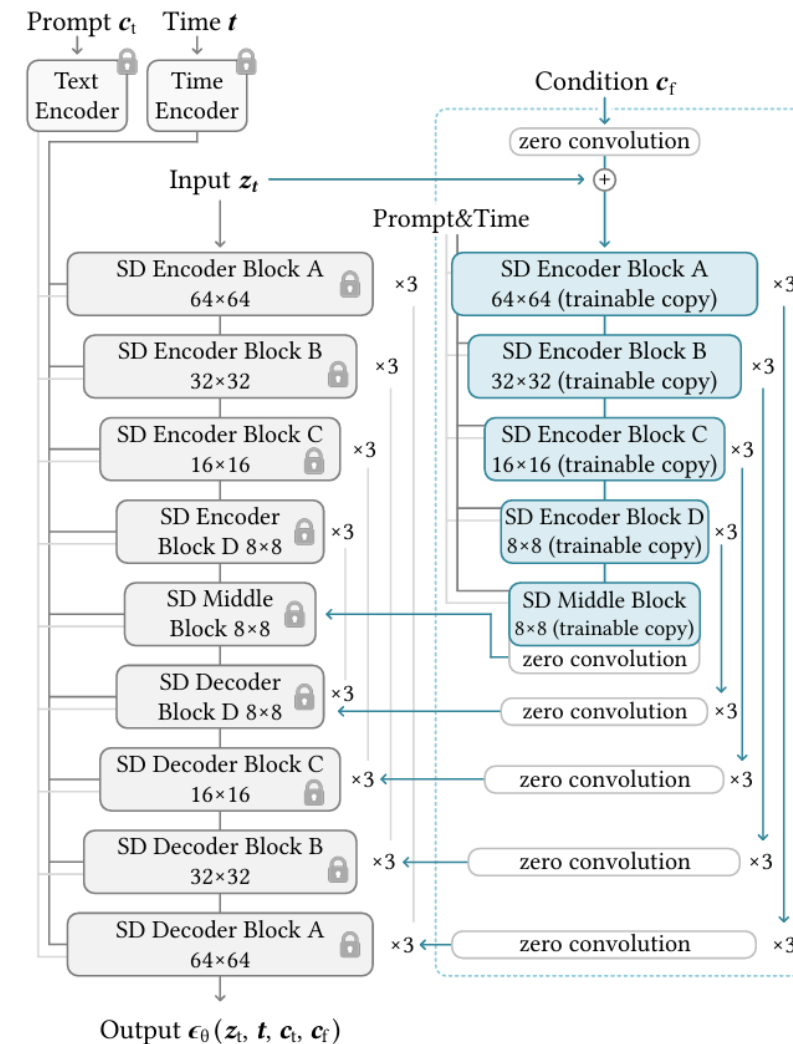
- Freeze the original NN block and make a trainable copy
- Add zero convolution layers (weights are zero)
- Zero convolution layer weights eventually become non-zero



Over time, the conditioning branch learns how much of the conditioning signal to inject!

# ControlNet with Stable Diffusion

- Augment encoder blocks and middle block
- Efficient: locked copy parameters are frozen
- Convert conditioning images to feature space vector matching Stable Diffusion size



(a) Stable Diffusion

(b) ControlNet

# Training

- Follow standard diffusion training and predict the noise added to a noisy image

$$\mathcal{L} = \mathbb{E}_{z_0, t, c_t, c_f, \epsilon \sim \mathcal{N}(0,1)} \left[ \|\epsilon - \epsilon_\theta(z_t, t, c_t, c_f)\|_2^2 \right] \quad (1)$$

- Randomly replace 50% of text prompts
- Zero convolutions add no additional noise, so image fidelity is preserved

# Training

- Follow standard diffusion training and predict the noise added to a noisy image

$$\mathcal{L} = \mathbb{E}_{z_0, t, c_t, c_f, \epsilon \sim \mathcal{N}(0,1)} \left[ \|\epsilon - \epsilon_\theta(z_t, t, c_t, c_f)\|_2^2 \right] \quad (1)$$

- Randomly replace 50% of text prompts
- Zero convolutions add no additional noise, so image fidelity is preserved

## "Sudden Convergence Phenomenon"



Test input



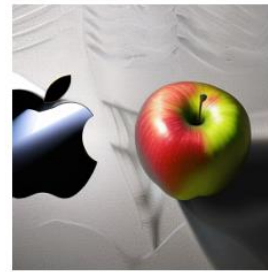
training step 100



step 1000



step 2000



step 6100



**step 6133**



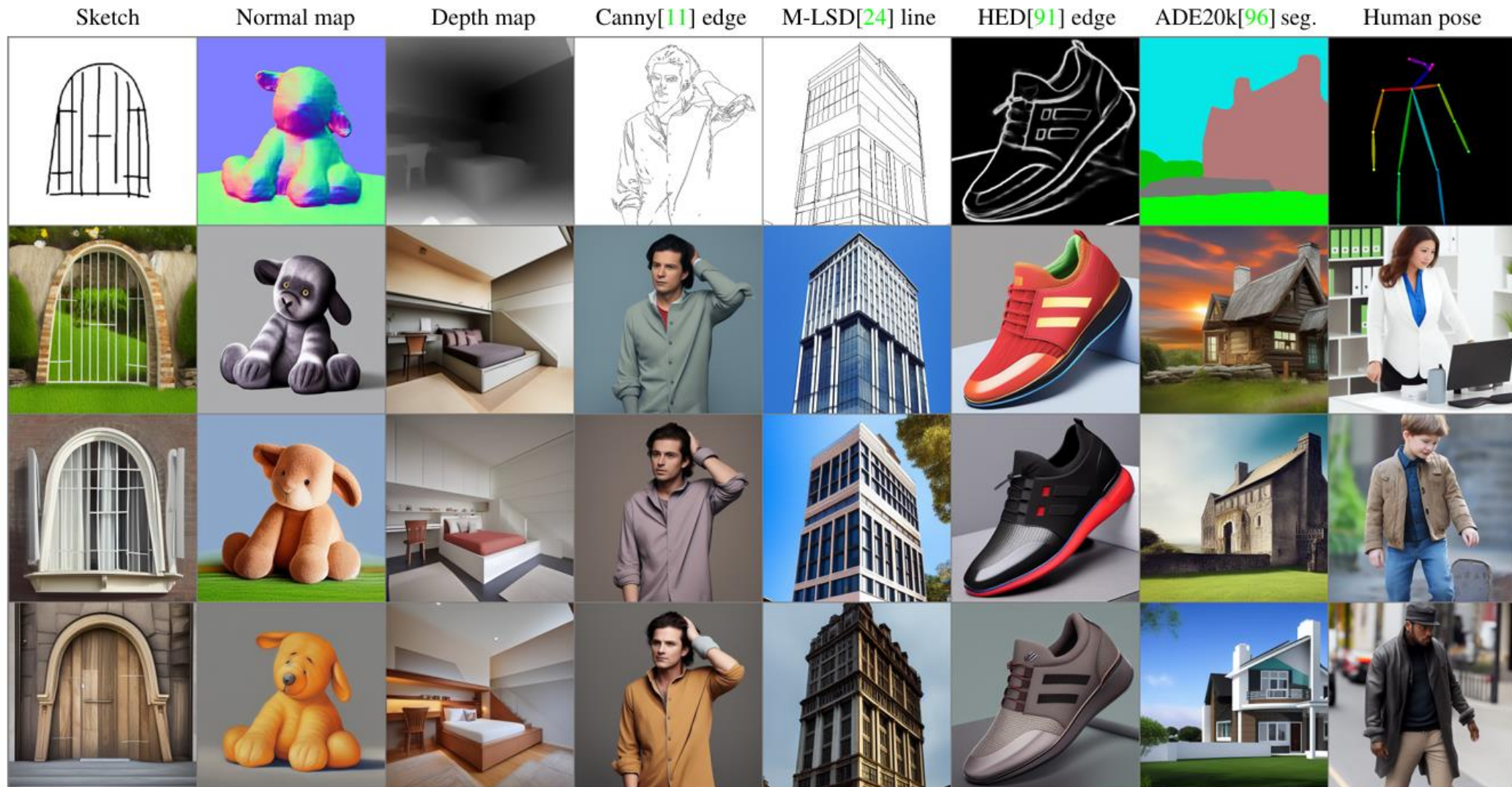
step 8000



step 12000

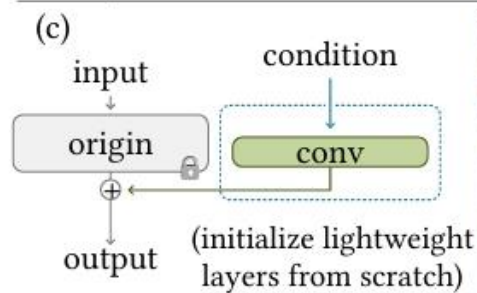
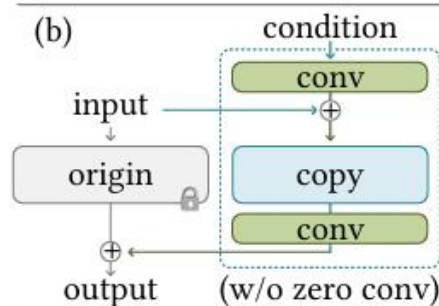
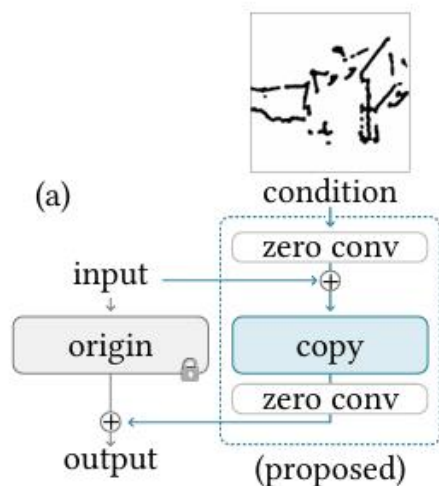


# Qualitative Results: No Prompts





# Ablations



No prompt



Insufficient prompt  
(w/o mentioning "house")  
*"high-quality and detailed masterpiece"*



Conflicting prompt  
*"delicious cake"*

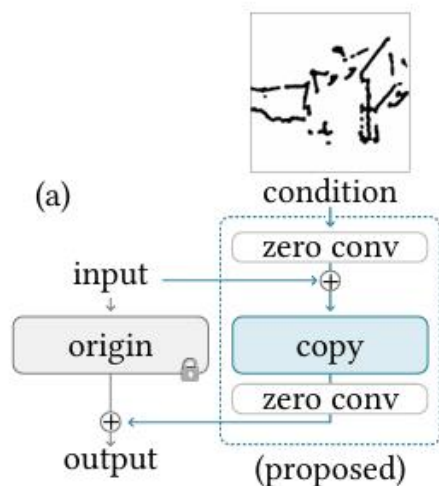


Perfect prompt  
*"a house, high-quality,  
extremely detailed, 4K, HQ"*





# Ablations



No prompt



Insufficient prompt  
(w/o mentioning "house")

*"high-quality and detailed masterpiece"*

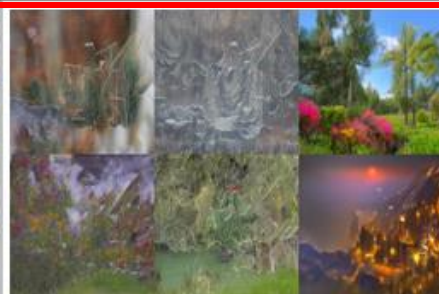
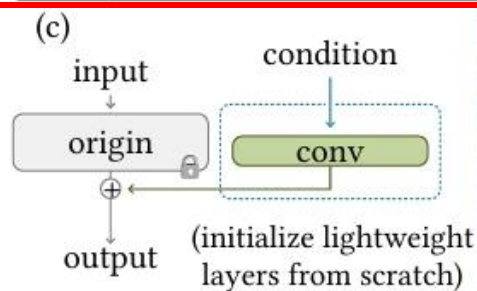
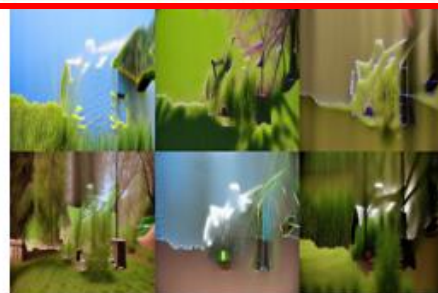
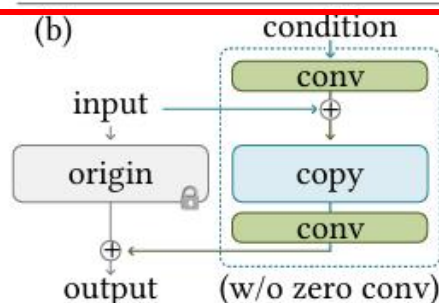


Conflicting prompt

*"delicious cake"*



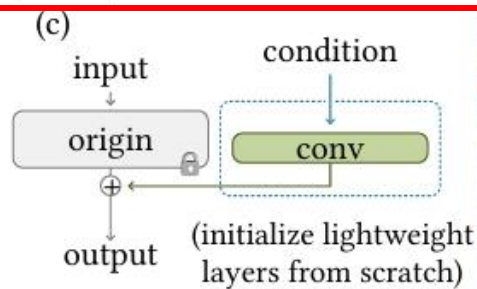
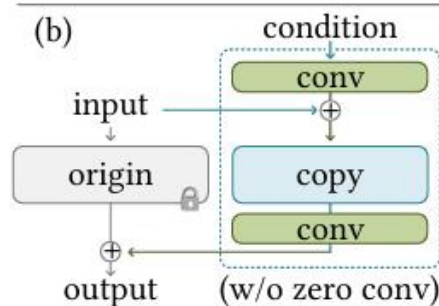
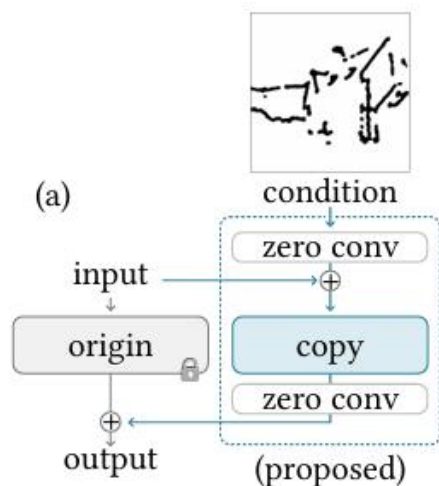
Perfect prompt  
*"a house, high-quality,  
extremely detailed, 4K, HQ"*





# Ablations

## ControlNet-lite



No prompt



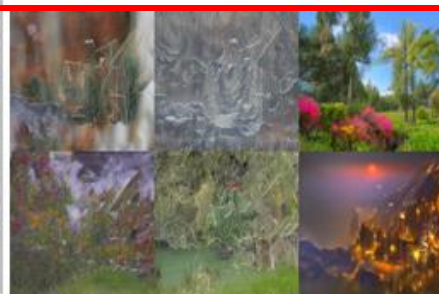
Insufficient prompt  
(w/o mentioning "house")  
"high-quality and detailed masterpiece"



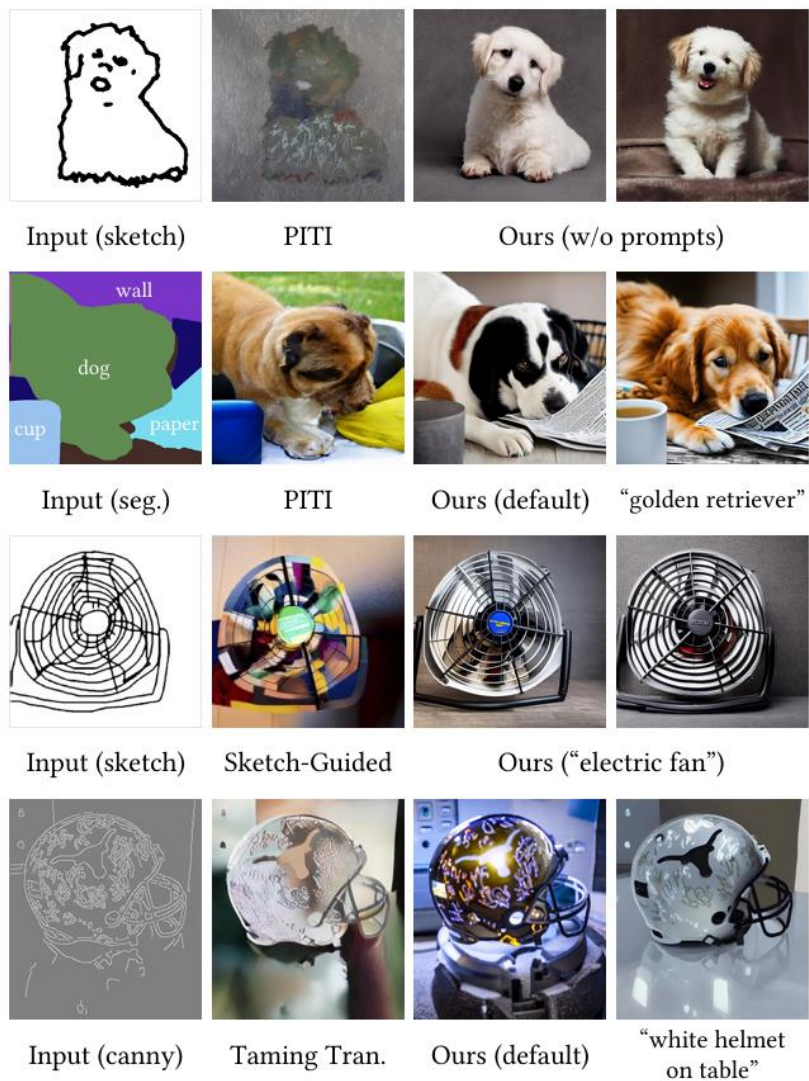
Conflicting prompt  
"delicious cake"



Perfect prompt  
"a house, high-quality,  
extremely detailed, 4K, HQ"



# Comparisons



Method	Result Quality $\uparrow$	Condition Fidelity $\uparrow$
PITI [89](sketch)	$1.10 \pm 0.05$	$1.02 \pm 0.01$
Sketch-Guided [88] ( $\beta = 1.6$ )	$3.21 \pm 0.62$	$2.31 \pm 0.57$
Sketch-Guided [88] ( $\beta = 3.2$ )	$2.52 \pm 0.44$	$3.28 \pm 0.72$
ControlNet-lite	$3.93 \pm 0.59$	$4.09 \pm 0.46$
ControlNet	<b><math>4.22 \pm 0.43</math></b>	<b><math>4.28 \pm 0.45</math></b>

Table 1: Average User Ranking (AUR) of result quality and condition fidelity. We report the user preference ranking (1 to 5 indicates worst to best) of different methods.



# Diversity

Method	FID ↓	CLIP-score ↑	CLIP-aes. ↑
Stable Diffusion	6.09	0.26	6.32
VQGAN [19](seg.)*	26.28	0.17	5.14
LDM [72](seg.)*	25.35	0.18	5.15
PITI [89](seg.)	19.74	0.20	5.77
ControlNet-lite	17.92	0.26	6.30
ControlNet	15.27	0.26	6.31

Table 3: Evaluation for image generation conditioned by semantic segmentation. We report FID, CLIP text-image score, and CLIP aesthetic scores for our method and other baselines. We also report the performance of Stable Diffusion without segmentation conditions. Methods marked with “\*” are trained from scratch.

**Thank You!**