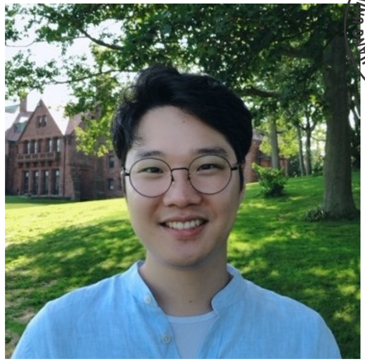
CAS: A Probability-Based Approach for Universal Condition Alignment Score

Chunsan Hong* [1]

Byunghee Cha* [2]



Tae-Hyun Oh [3]

[1]Korea Advanced Institute of Science & Technology (KAIST)

[2]Seoul National University

[3]Pohang University of Science & Technology (POSTECH)

* Equal contribution

Diffusion models can be designed with various conditions and outputs

Text-to-Image [1]   {Edge/Text}-to-Image [2]   Text-to-Audio [3]

[1] Romach et. al., High-Resolution Image Synthesis with Latent Diffusion Models
[2] Zhang et. al., Adding Conditional Control to Text-to-Image Diffusion Models
[3] Liu et. al., AudioLDM: Text-to-Audio Generation with Latent Diffusion Models

Diffusion models can be designed with various conditions and outputs

Text ➡ 

Text-to-Image [1]

\+ Text ➡

{Edge/Text}-to-Image [2]

Text ➡

Text-to-Audio [3]

[1] Romach et. al., High-Resolution Image Synthesis with Latent Diffusion Models
[2] Zhang et. al., Adding Conditional Control to Text-to-Image Diffusion Models
[3] Liu et. al., AudioLDM: Text-to-Audio Generation with Latent Diffusion Models

Diffusion models can be designed with various conditions and outputs

Text ➡ 

Text-to-Image [1]

\+ Text ➡ 

{Edge/Text}-to-Image [2]

Text ➡

Text-to-Audio [3]

[1] Romach et. al., High-Resolution Image Synthesis with Latent Diffusion Models
[2] Zhang et. al., Adding Conditional Control to Text-to-Image Diffusion Models
[3] Liu et. al., AudioLDM: Text-to-Audio Generation with Latent Diffusion Models

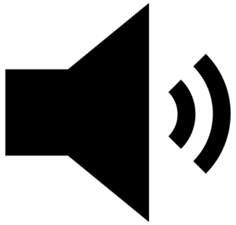
Diffusion models can be designed with various conditions and outputs

Text ➡ 

Text-to-Image [1]

\+ Text ➡ 

{Edge/Text}-to-Image [2]

Text ➡ 

Text-to-Audio [3]

[1] Romach et. al., High-Resolution Image Synthesis with Latent Diffusion Models
[2] Zhang et. al., Adding Conditional Control to Text-to-Image Diffusion Models
[3] Liu et. al., AudioLDM: Text-to-Audio Generation with Latent Diffusion Models

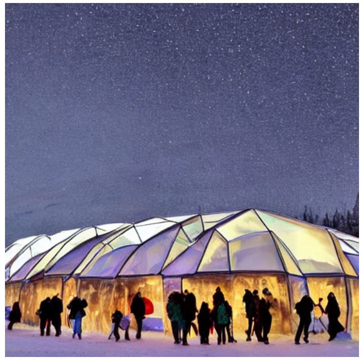
Problem in Conditional Diffusion Models

Generated samples **do not always follow** the conditions

"People …,
Aurora …,
Igloos …"

Conditional Diffusion Model

Generate



Aurora?

Igloo, Aurora?

People, Aurora?

Problem in Conditional Diffusion Models

Generated samples **do not always follow** the conditions

What people do? **Cherry Picking** by **hands!**

“People …, Aurora …, Igloos …”

Cherry Pick

Good!

No Aurora

No Igloo
No Aurora

No People
No Aurora

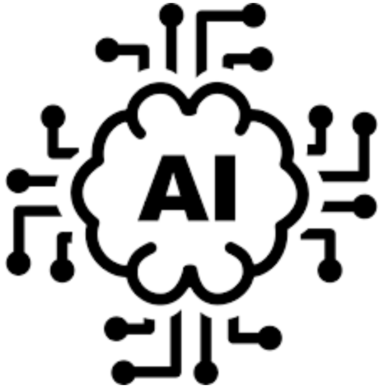
Problem in Conditional Diffusion Models

Generated samples **do not always follow** the conditions

➡ What people do? **Cherry Picking** by **hands!**

➡ **Automatic Scoring** via AI is **required**

“People …, Aurora …, Igloos …” ➡



➡ **Automated Scoring**

Previous Work

Common framework: Data Construction → Train [1, 2, 3]

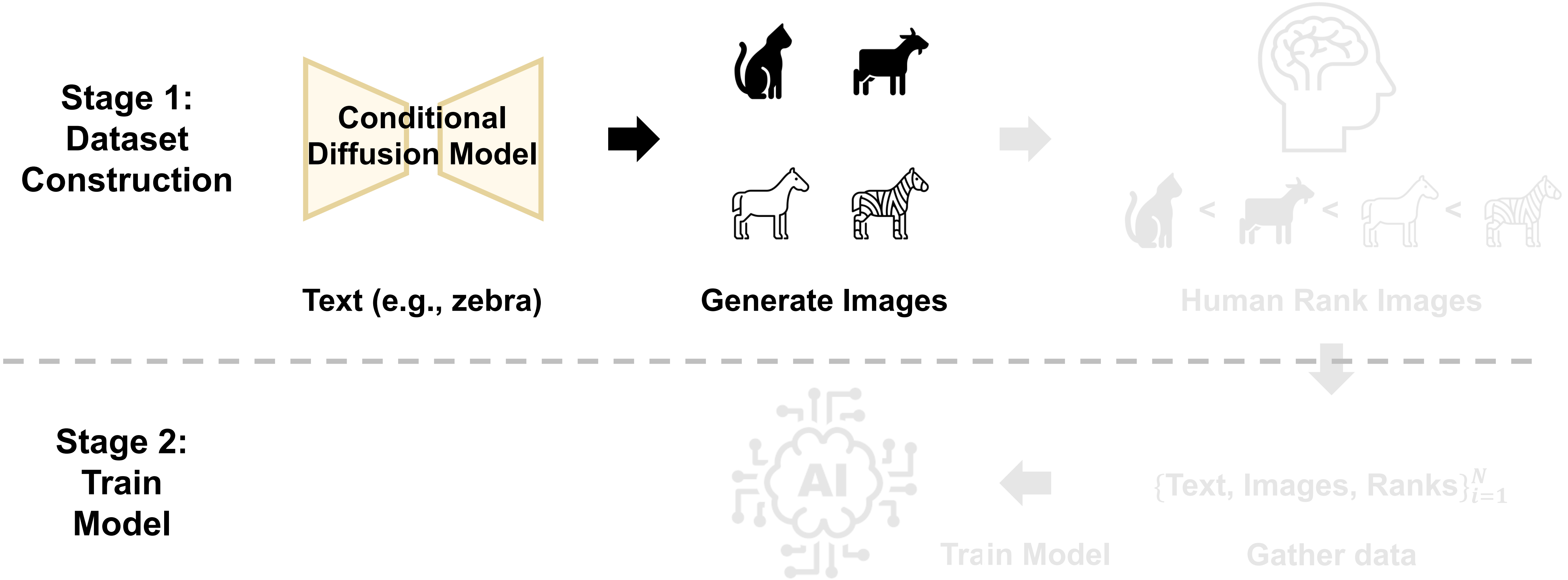
Stage 1: Dataset Construction

Stage 2: Train Model

[1] Kirstain et. al., Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation
[2] Su et. al., ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation
[3] Wu et. al., Human Preference Score: Better Aligning Text-to-Image Models with Human Preference

Previous Work

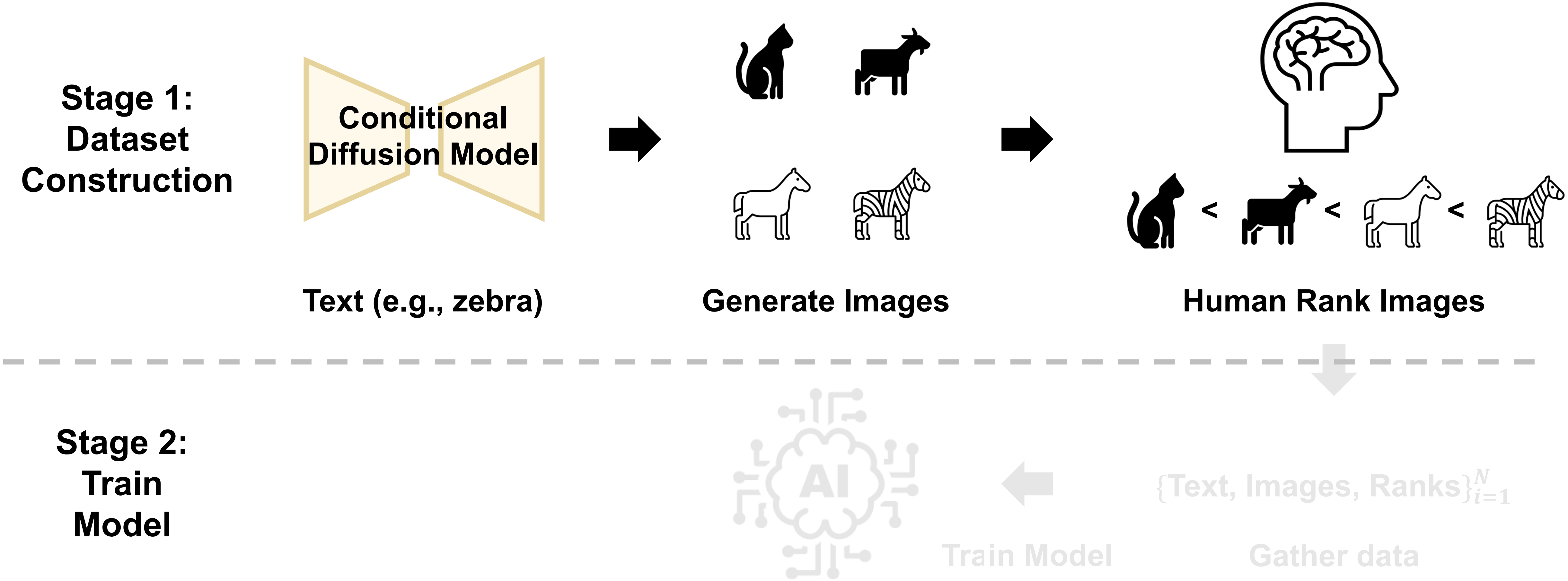
Common framework: Data Construction → Train [1, 2, 3]



[1] Kirstain et. al., Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation
[2] Su et. al., ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation
[3] Wu et. al., Human Preference Score: Better Aligning Text-to-Image Models with Human Preference

Previous Work

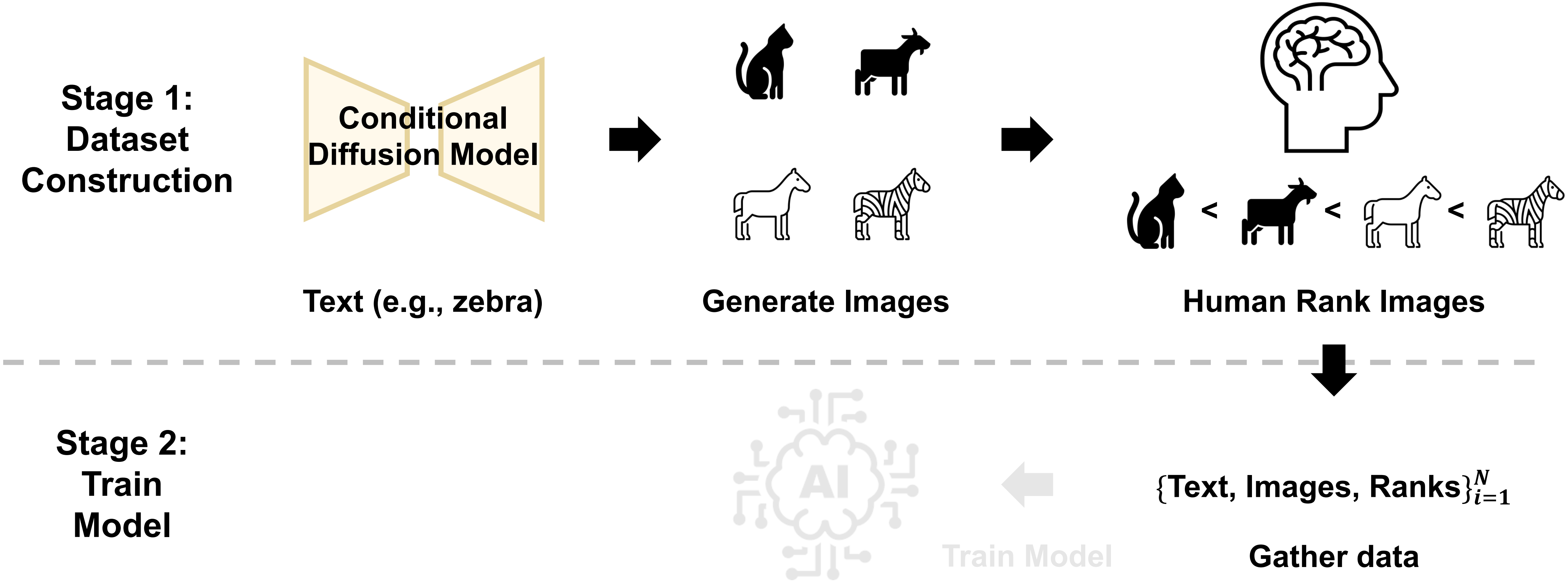
Common framework: Data Construction → Train [1, 2, 3]



[1] Kirstain et. al., Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation
[2] Su et. al., ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation
[3] Wu et. al., Human Preference Score: Better Aligning Text-to-Image Models with Human Preference

Previous Work

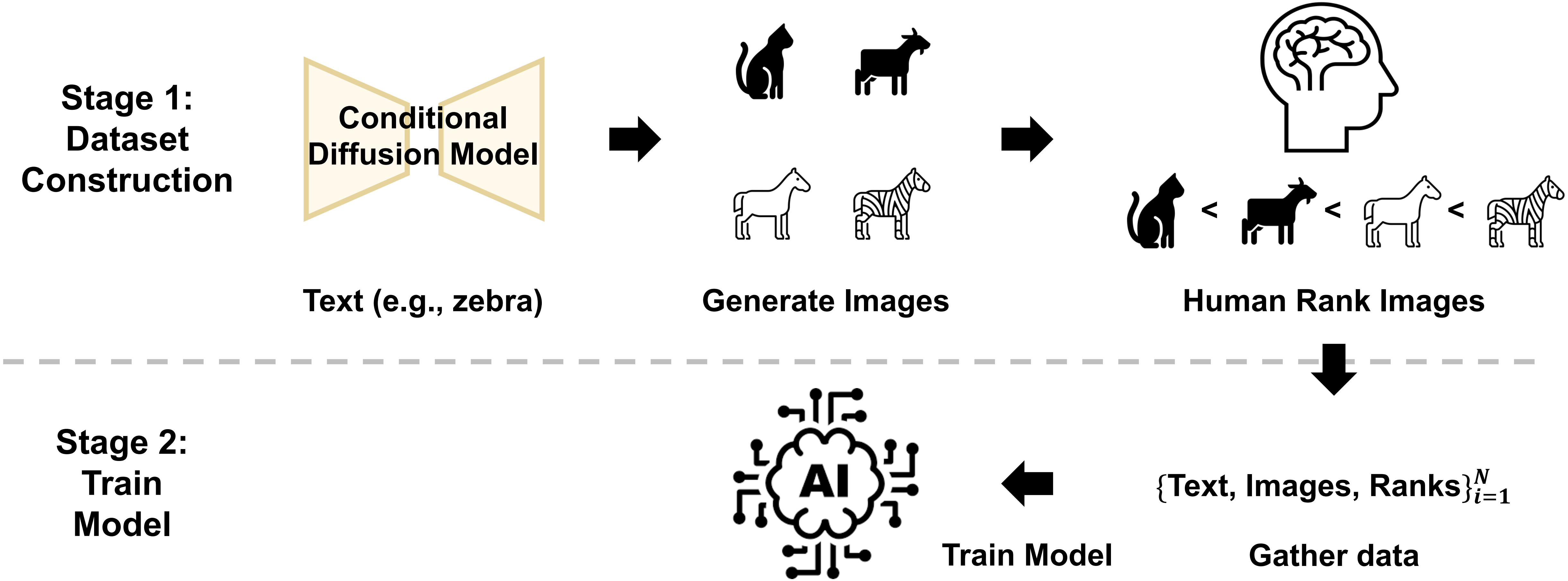
Common framework: Data Construction → Train [1, 2, 3]



[1] Kirstain et. al., Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation
[2] Su et. al., ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation
[3] Wu et. al., Human Preference Score: Better Aligning Text-to-Image Models with Human Preference

Previous Work

Common framework: Data Construction → Train [1, 2, 3]



[1] Kirstain et. al., Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation
[2] Su et. al., ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation
[3] Wu et. al., Human Preference Score: Better Aligning Text-to-Image Models with Human Preference

Previous Work

Common framework: Data Construction → Train [1, 2, 3]

Stage 1: Dataset Construction

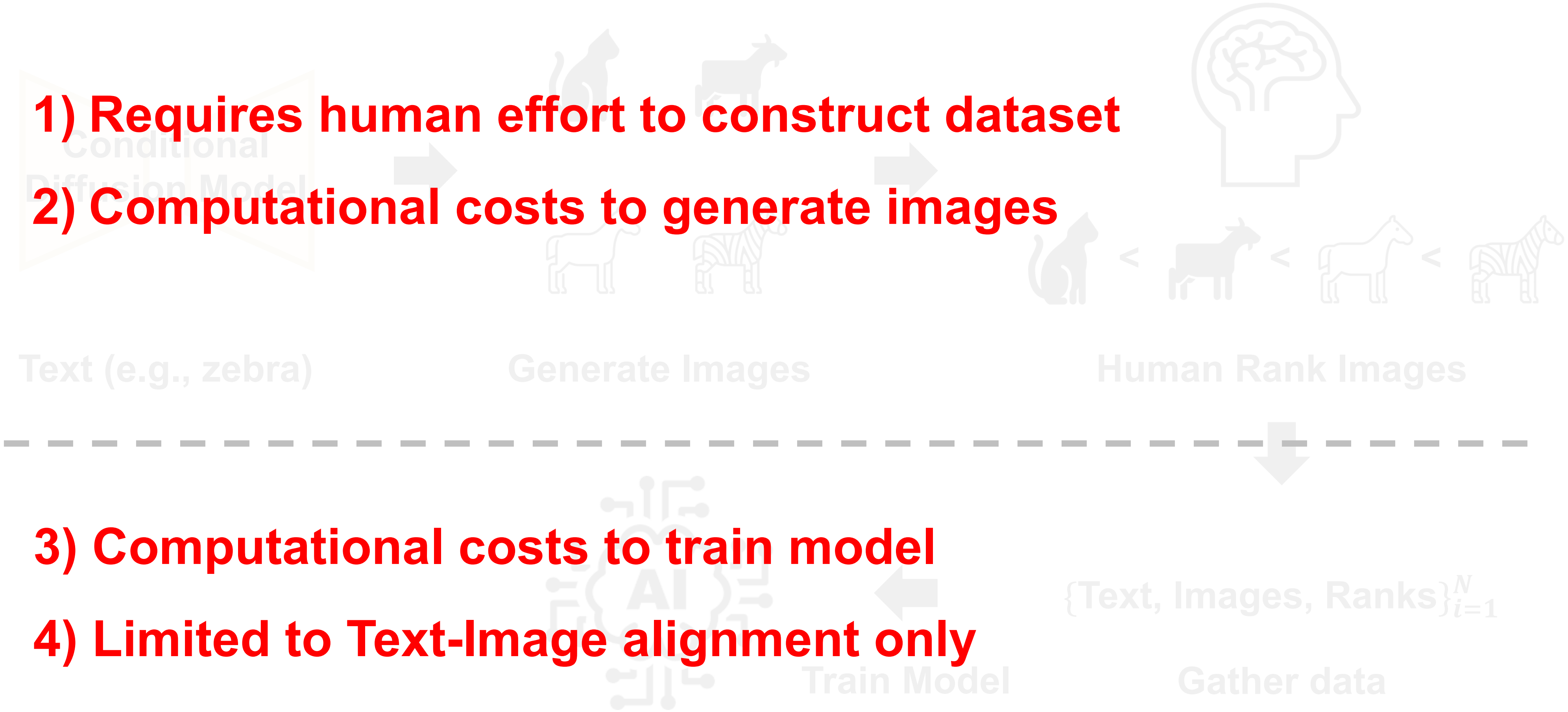
1) Requires human effort to construct dataset

2) Computational costs to generate images

Stage 2: Train Model

3) Computational costs to train model

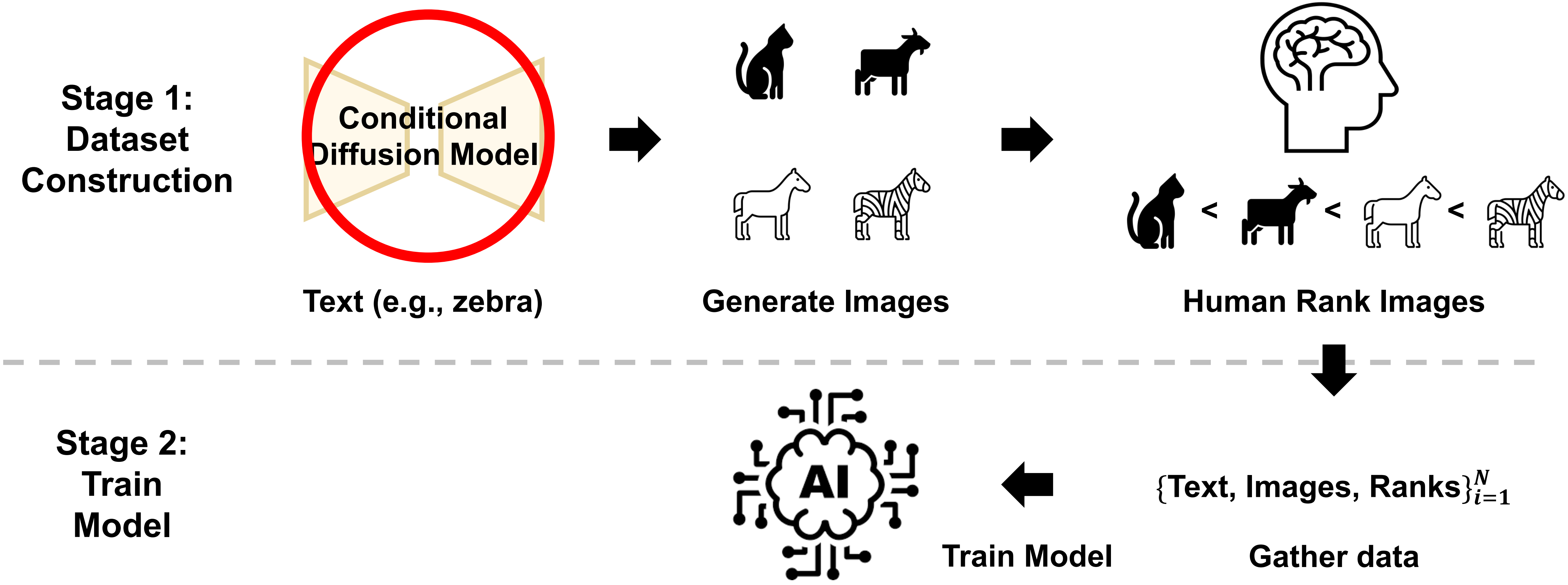
4) Limited to Text-Image alignment only



[1] Kirstain et. al., Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation
[2] Su et. al., ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation
[3] Wu et. al., Human Preference Score: Better Aligning Text-to-Image Models with Human Preference

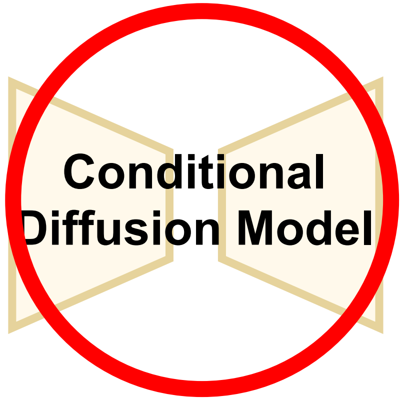
Previous Work

Common framework: Data Construction → Train [1, 2, 3]



[1] Kirstain et. al., Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation
[2] Su et. al., ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation
[3] Wu et. al., Human Preference Score: Better Aligning Text-to-Image Models with Human Preference

Our Key Idea



Conditional Diffusion Models might already know alignments between condition and generated samples

Use the Conditional Probability measured by the diffusion models themselves

1) **Training Free**
2) **Can be applied to any type of {condition/output}**

e.g.) {text/image}, {image/image}, {text/audio}

Preliminary

Score-Based Generative Modeling through Stochastic Differential Equations (Song et. al, ICLR 2021)

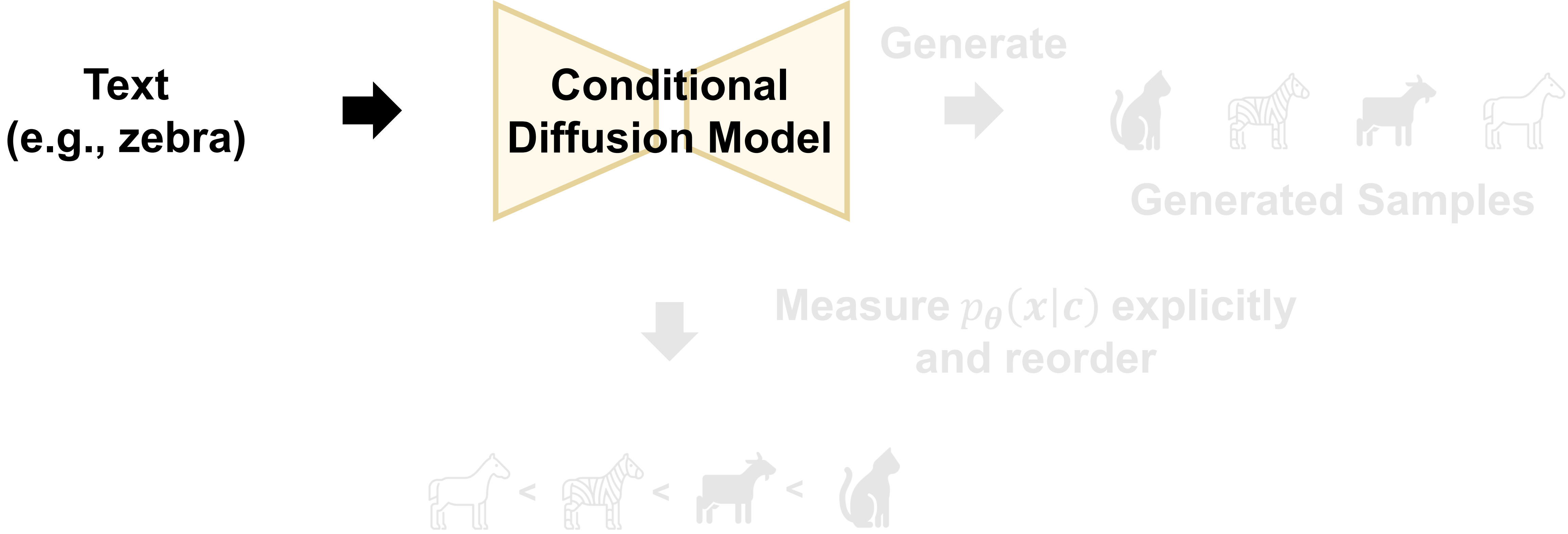
If the reverse process of a conditional diffusion model is formulated as a probability flow ODE, it can be represented by the following equation:

$$d\boldsymbol{x} = \boldsymbol{f}_\theta(\boldsymbol{x}(t), c, t)dt.$$

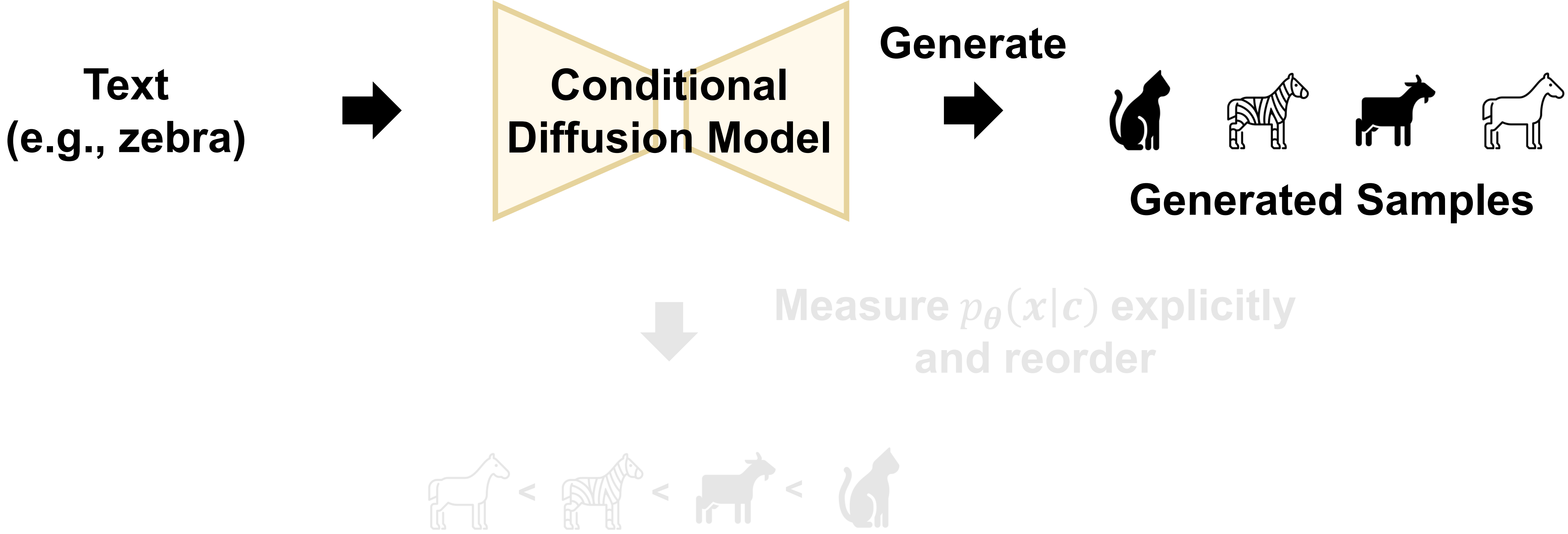
Then, **conditional probability** can be measured as follows:

$$\log p_0(\boldsymbol{x}(0)|c) = \log p_1\big(\boldsymbol{x}(1)\big) + \int_0^1 \nabla_{\boldsymbol{x}} \cdot \boldsymbol{f}_\theta(\boldsymbol{x}(t), c, t)dt\,.$$

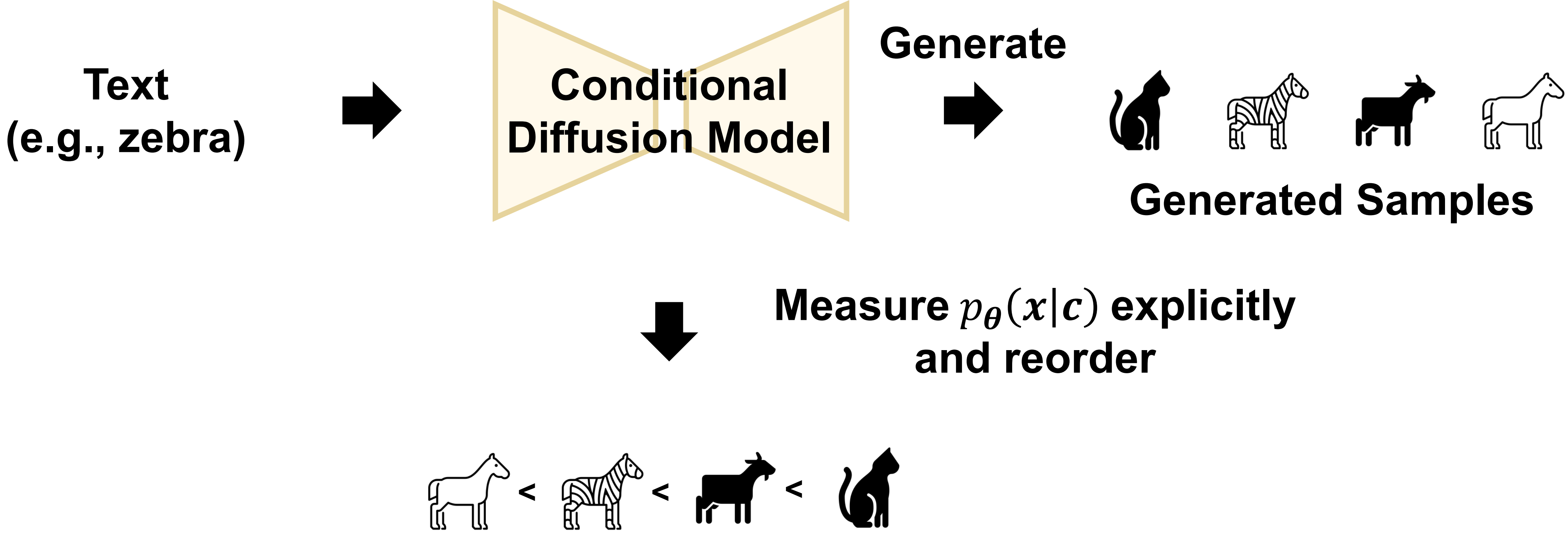
Our Initial attempt



Our Initial attempt



Our Initial attempt

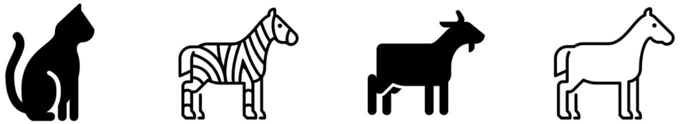


Our Initial attempt

Text
(e.g., zebra)

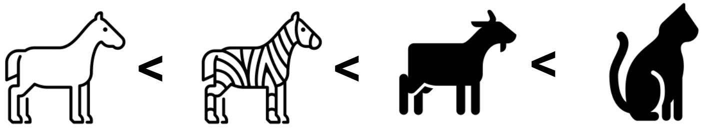
Conditional
Diffusion Model

Generate



Generated Samples

Measure $p_\theta(x|c)$ explicitly
and reorder



3rd

1st

Not working.
Why?

Preliminary experiment

Goal: Find other options for alignment score

How to: Find option ∝ CLIP Score

Experiment step:

1. Generate 100 images from text "Woman, Green hair, Sunglasses"
2. Measure
   - $\log p_\theta(x)$
   - $\log p_\theta(x|c)$
   - $\log p_\theta(x|c) - \log p_\theta(x)$
3. Select Top N% via each probability and measure mean CLIP score

Preliminary experiment

Goal: Find other options for alignment score

How to: Find option ∝ CLIP Score

Experiment step:

1. Generate 100 images from text "Woman, Green hair, Sunglasses"
2. Measure
   - $\log p_\theta (x)$
   - $\log p_\theta (x|c)$
   - $\log p_\theta (x|c) - \log p_\theta (x)$
3. Select Top N% via each probability and measure mean CLIP score

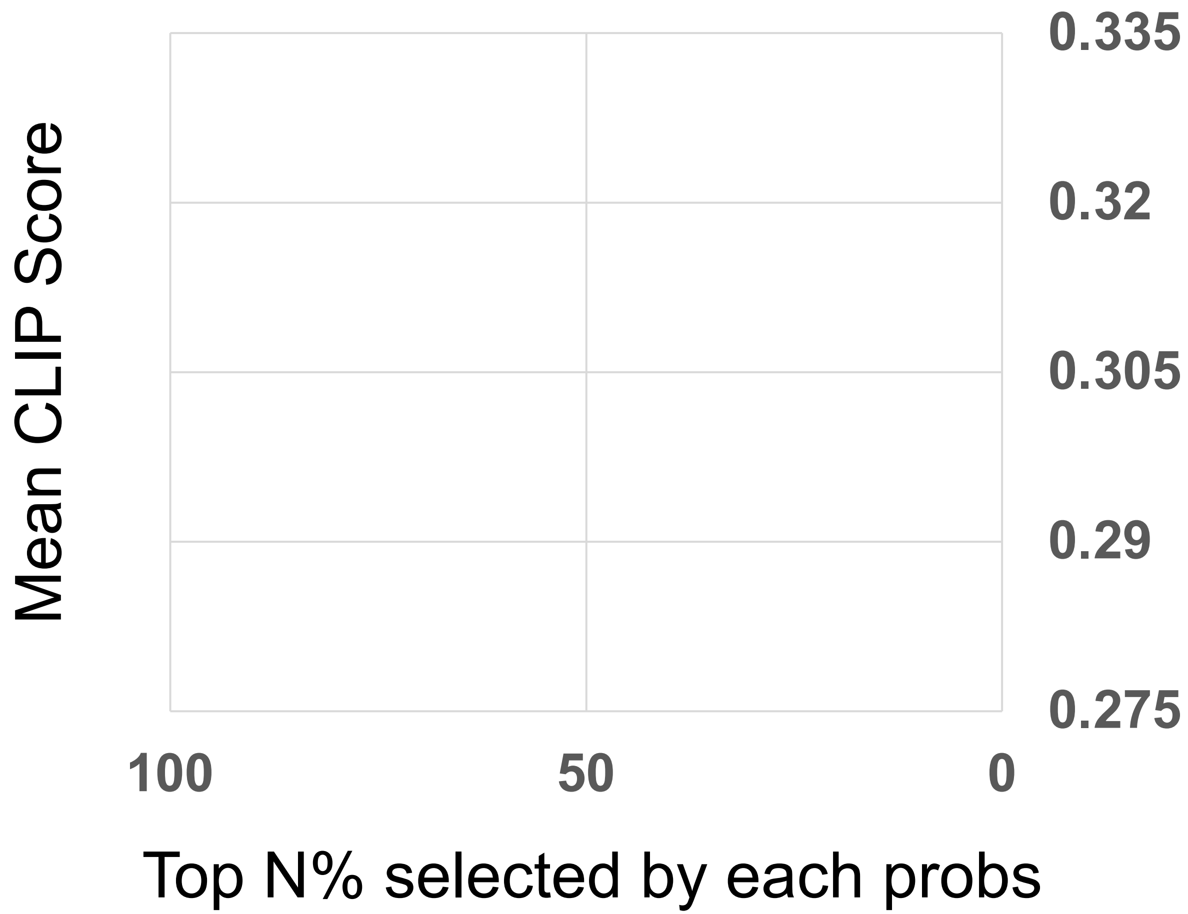
Preliminary experiment

Goal: Find other options for alignment score

How to: Find option ∝ CLIP Score

Experiment step:

1. Generate 100 images from text “Woman, Green hair, Sunglasses”
2. Measure
   - $\log p_\theta (x)$
   - $\log p_\theta (x|c)$
   - $\log p_\theta (x|c) - \log p_\theta (x)$
3. Select Top N% via each probability and measure mean CLIP score



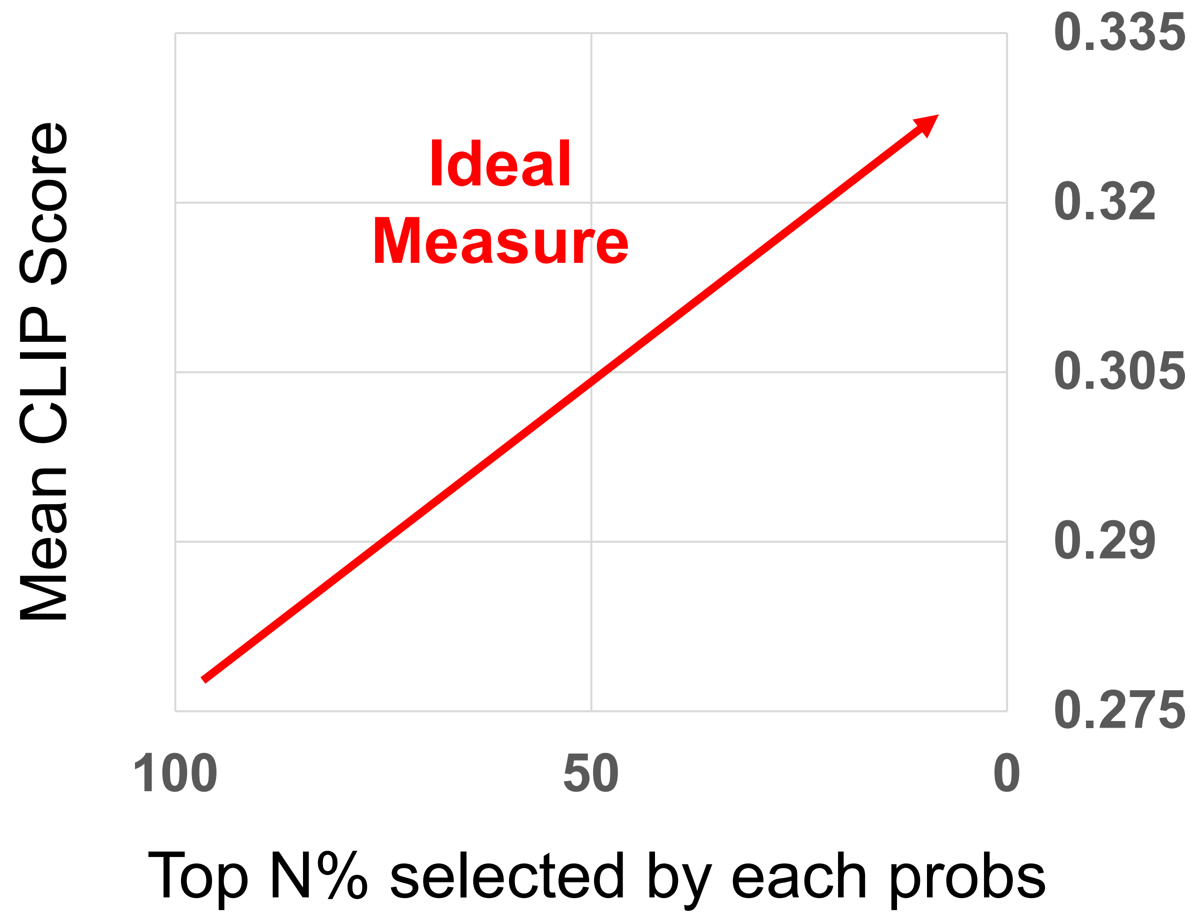
Preliminary experiment

Goal: Find other options for alignment score

How to: Find option ∝ CLIP Score

Experiment step:

1. Generate 100 images from text “Woman, Green hair, Sunglasses”
2. Measure
    - $\log p_\theta(x)$
    - $\log p_\theta(x|c)$
    - $\log p_\theta(x|c) - \log p_\theta(x)$
3. Select Top N% via each probability and measure mean CLIP score



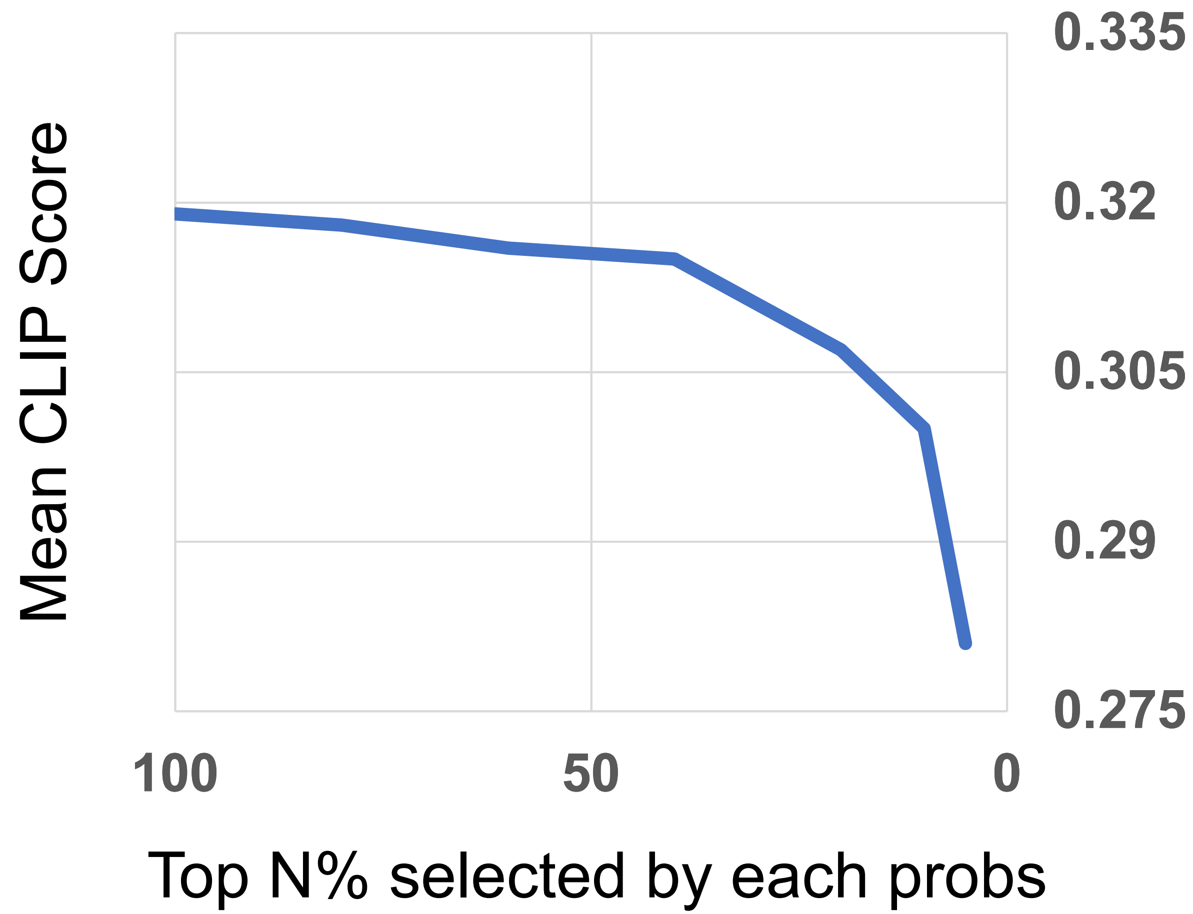
Preliminary experiment

Goal: Find other options for alignment score

How to: Find option ∝ CLIP Score

Experiment step:

1. Generate 100 images from text “Woman, Green hair, Sunglasses”
2. Measure
   - $\log p_\theta(x)$: Inverse proportional
   - $\log p_\theta(x|c)$
   - $\log p_\theta(x|c) - \log p_\theta(x)$
3. Select Top N% via each probability and measure mean CLIP score



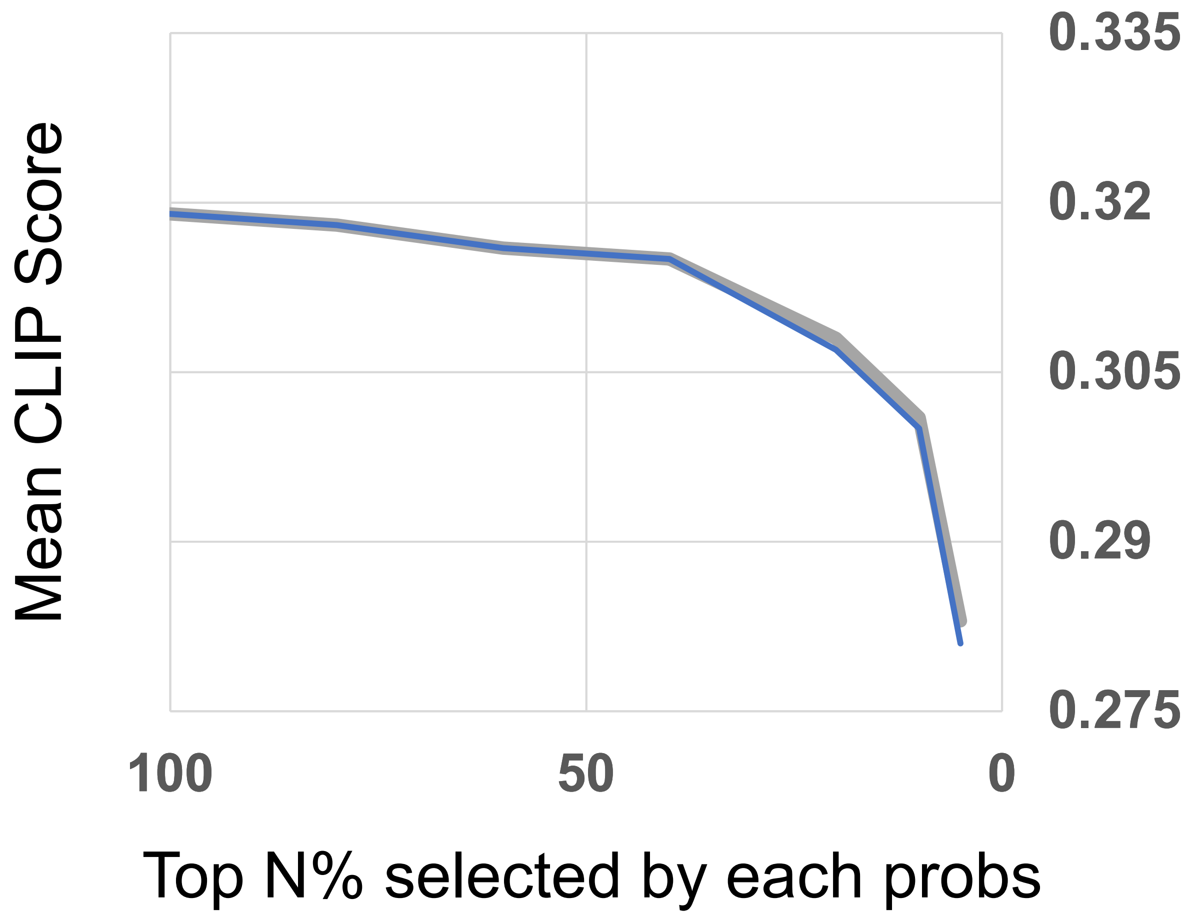
Preliminary experiment

Goal: Find other options for alignment score

How to: Find option ∝ CLIP Score

Experiment step:

1. Generate 100 images from text "Woman, Green hair, Sunglasses"
2. Measure
   - $\log p_\theta(x)$: Inverse proportional
   - $\log p_\theta(x|c)$: overfit to $p_\theta(x)$
   - $\log p_\theta(x|c) - \log p_\theta(x)$
3. Select Top N% via each probability and measure mean CLIP score



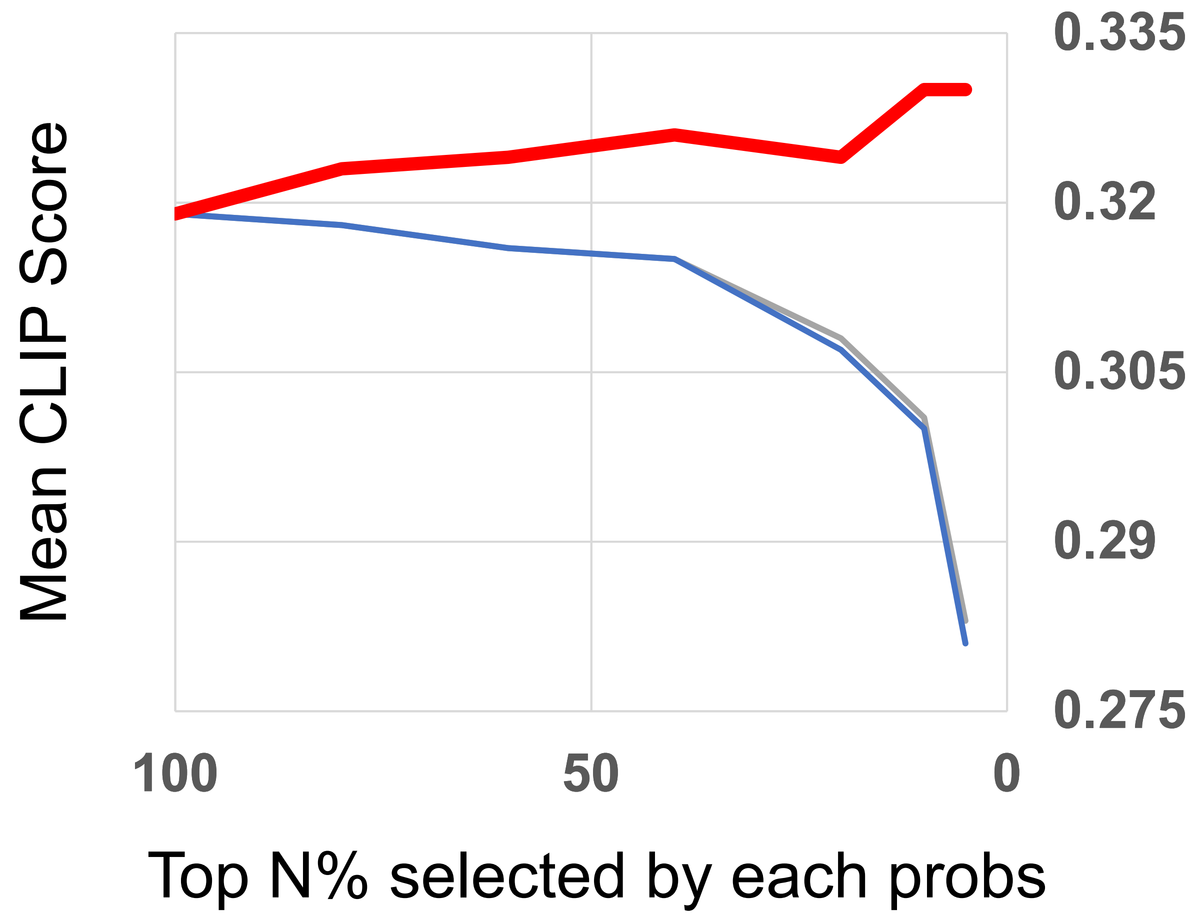
Preliminary experiment

Goal: Find other options for alignment score

How to: Find option ∝ CLIP Score

Experiment step:

1. Generate 100 images from text “Woman, Green hair, Sunglasses”
2. Measure
   - $\log p_\theta(x)$: Inverse proportional
   - $\log p_\theta(x|c)$: overfit to $p_\theta(x)$
   - $\log p_\theta(x|c) - \log p_\theta(x)$
3. Select Top N% via each probability and measure mean CLIP score



Insight from toy Experiment

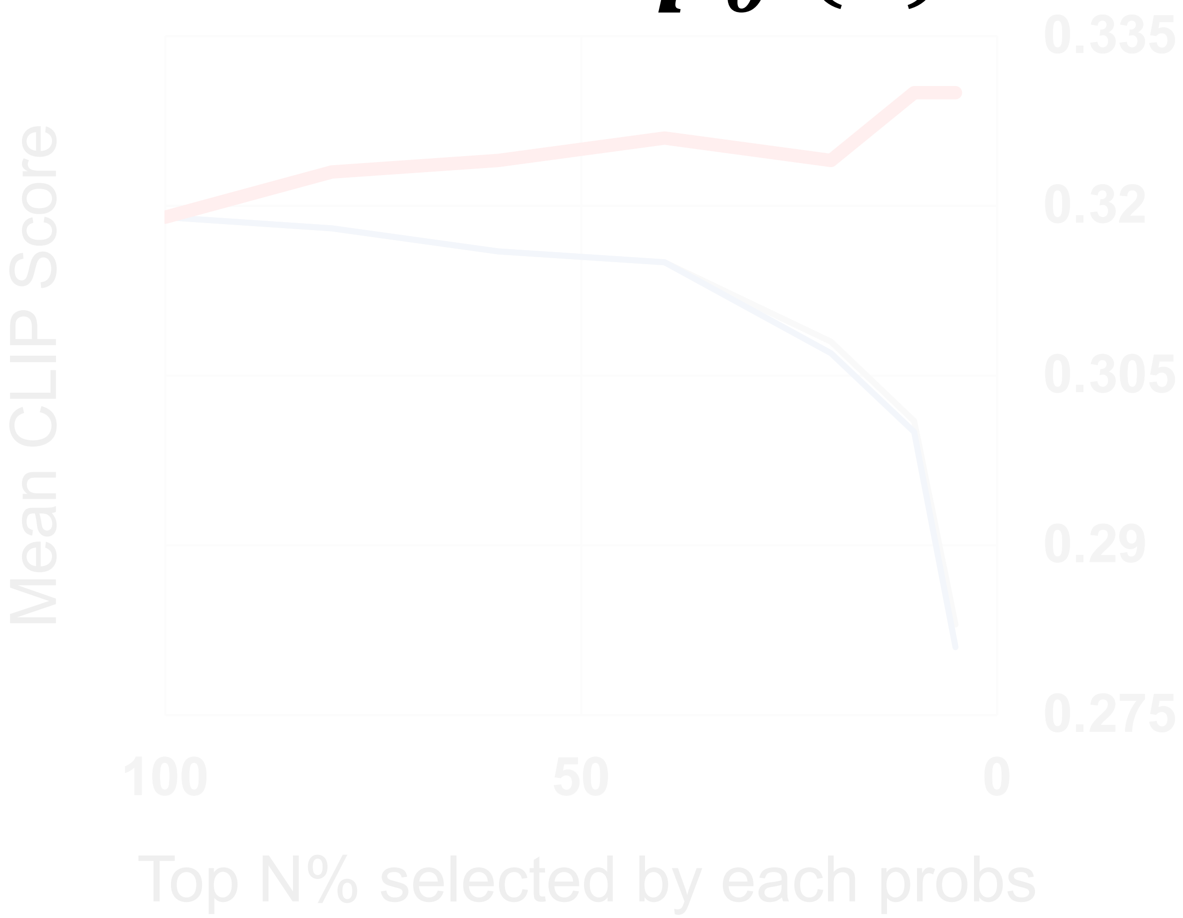
$\log p_\theta(x|c) - \log p_\theta(x)$ **is better than** $p_\theta(x)$.

Goal: Find other options for alignment score

How to: Find option ∝ CLIP Score

Experiment step:

1. Generate 100 images from text "Woman, Green hair, Sunglasses"
2. Measure
   - $\log p_\theta(x)$: Inverse proportional
   - $\log p_\theta(x|c)$: overfit to $p_\theta(x)$
   - $\log p_\theta(x|c) - \log p_\theta(x)$
3. Select Top N% via each probability and measure mean CLIP score



Insight from toy Experiment

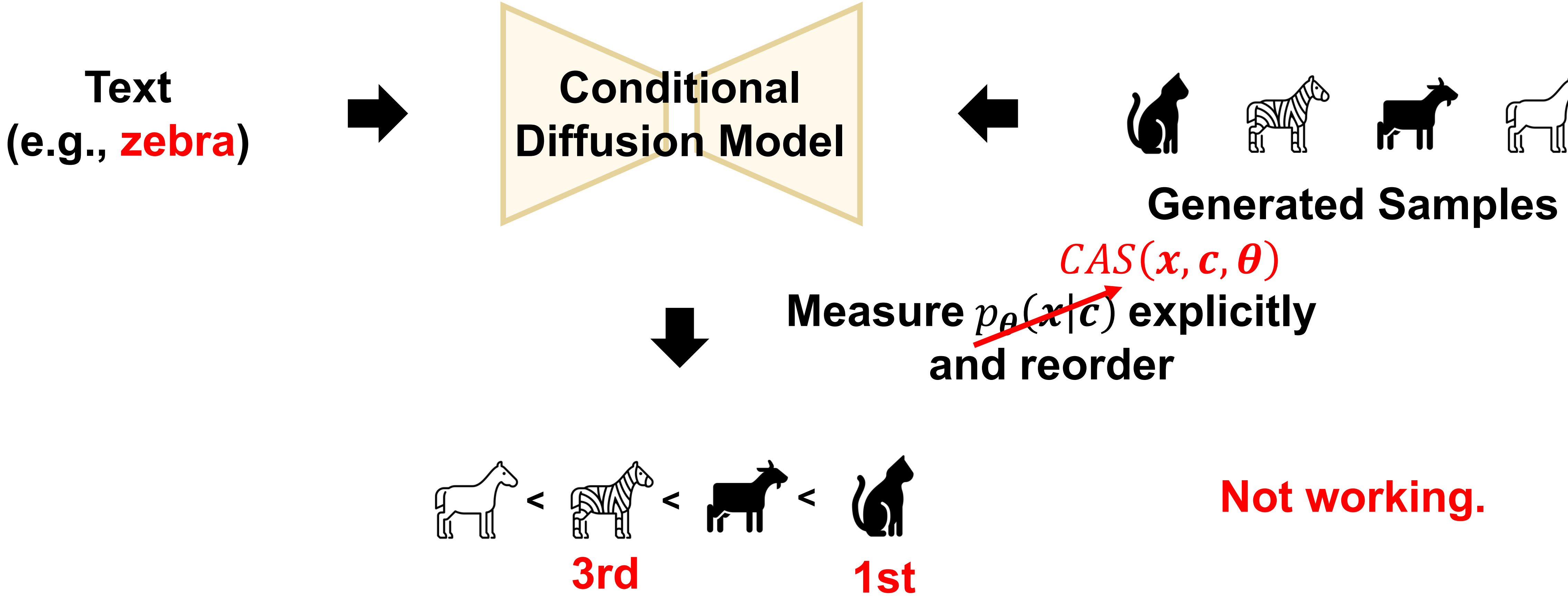
$\log p_\theta(x|c) - \log p_\theta(x)$ is better than $p_\theta(x)$.

Therefore,

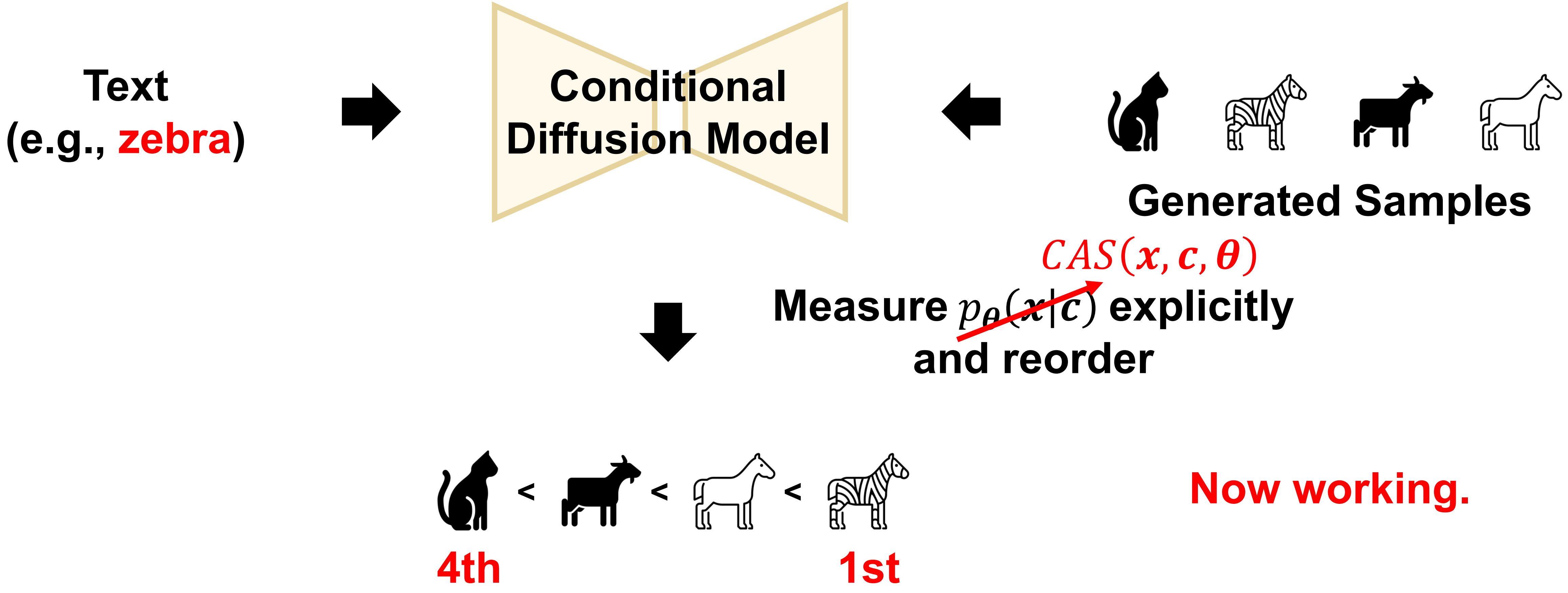
we define our universal Condition Alignment Score as:

$$CAS(x, c, \theta) = \log p_\theta(x|c) - \log p_\theta(x)$$

Our Method

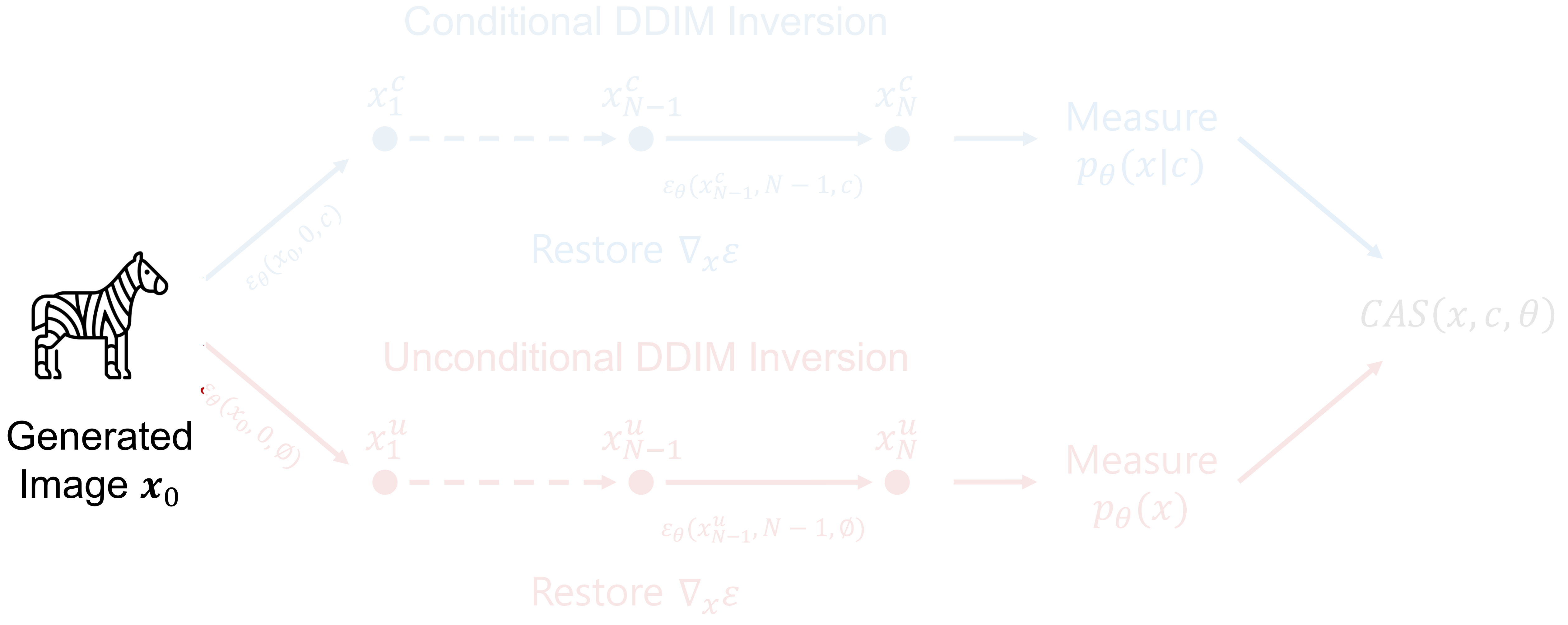

Text
(e.g., zebra)
Conditional Diffusion Model
Generated Samples
$CAS(x, c, \theta)$
Measure $\cancel{p_\theta(x|c)}$ explicitly
and reorder
< < <
3rd
1st
Not working.

Our Method

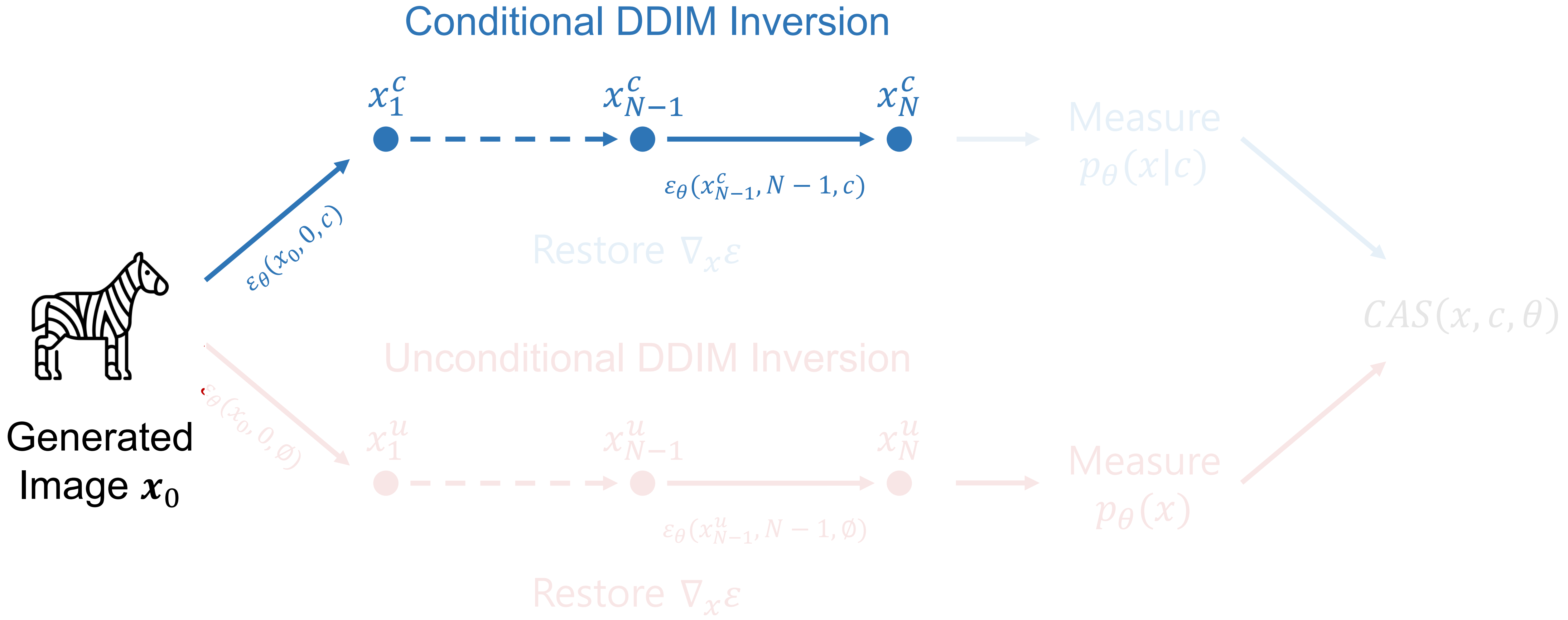


Now working.

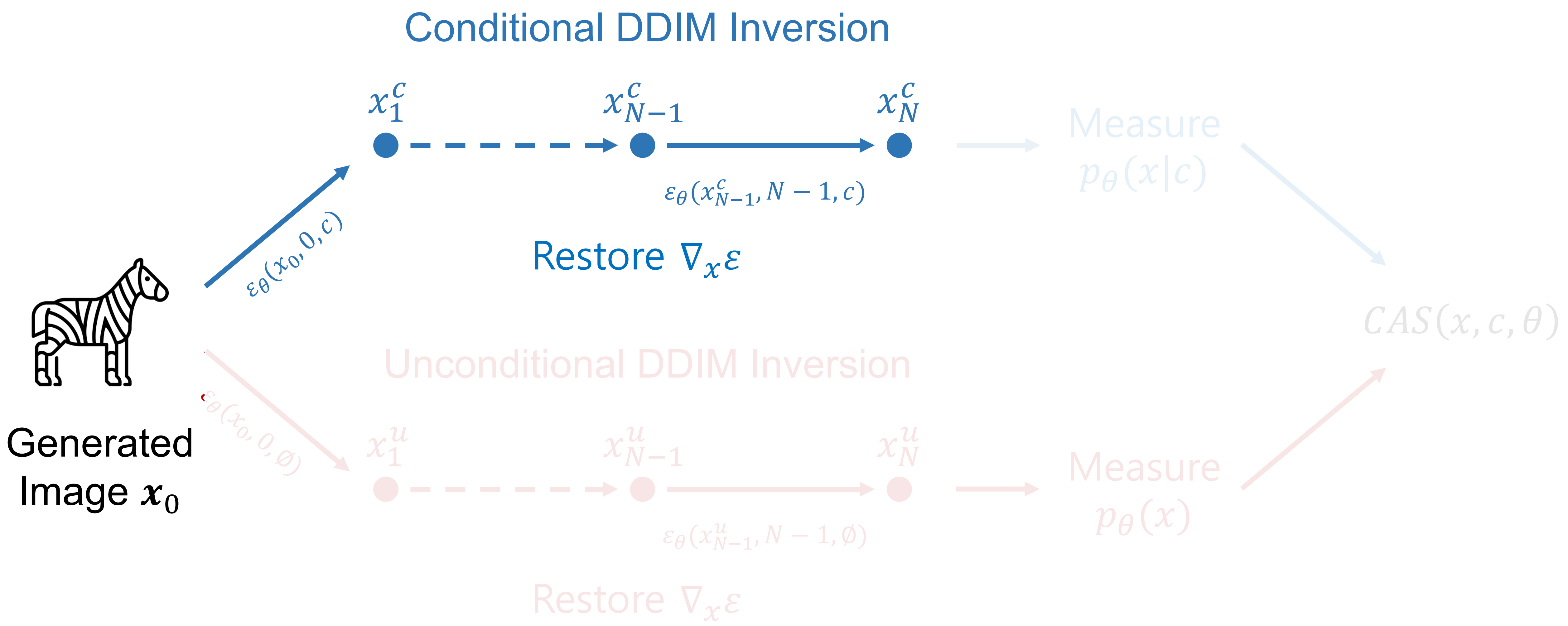
Method in detail:



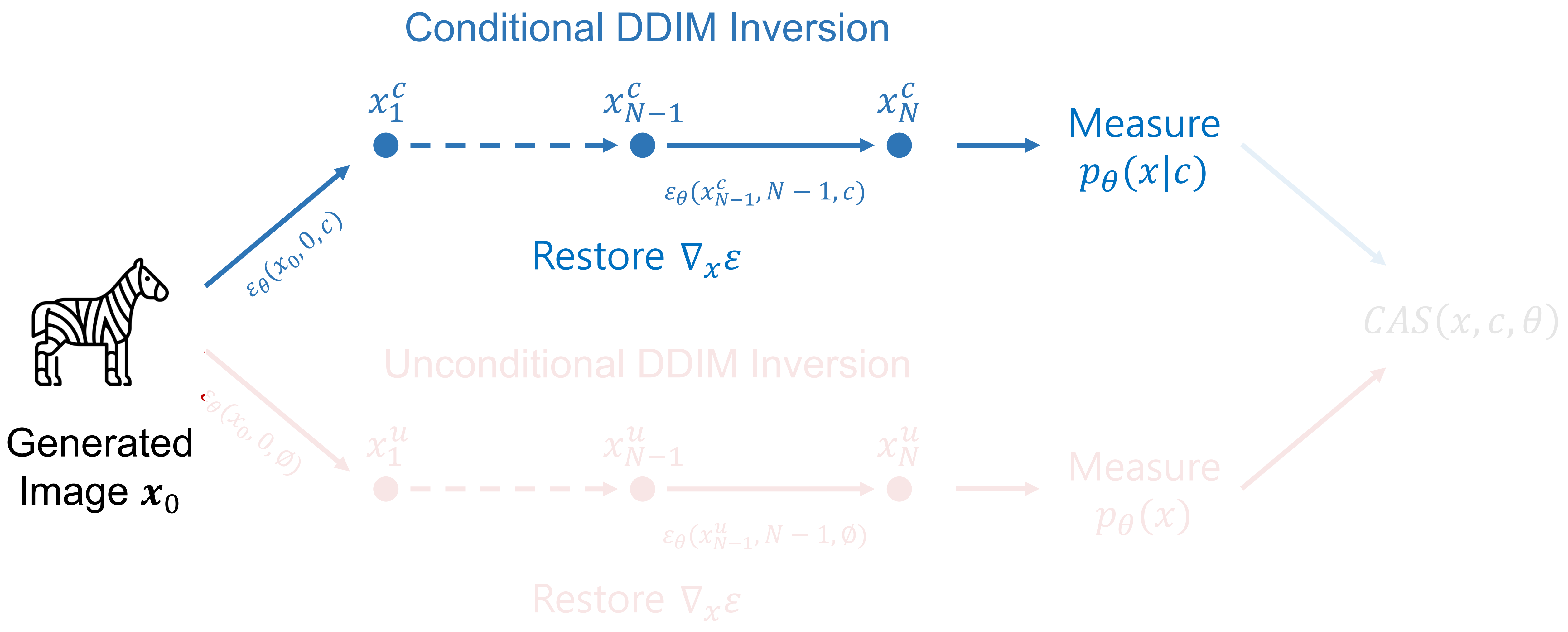
Method in detail:



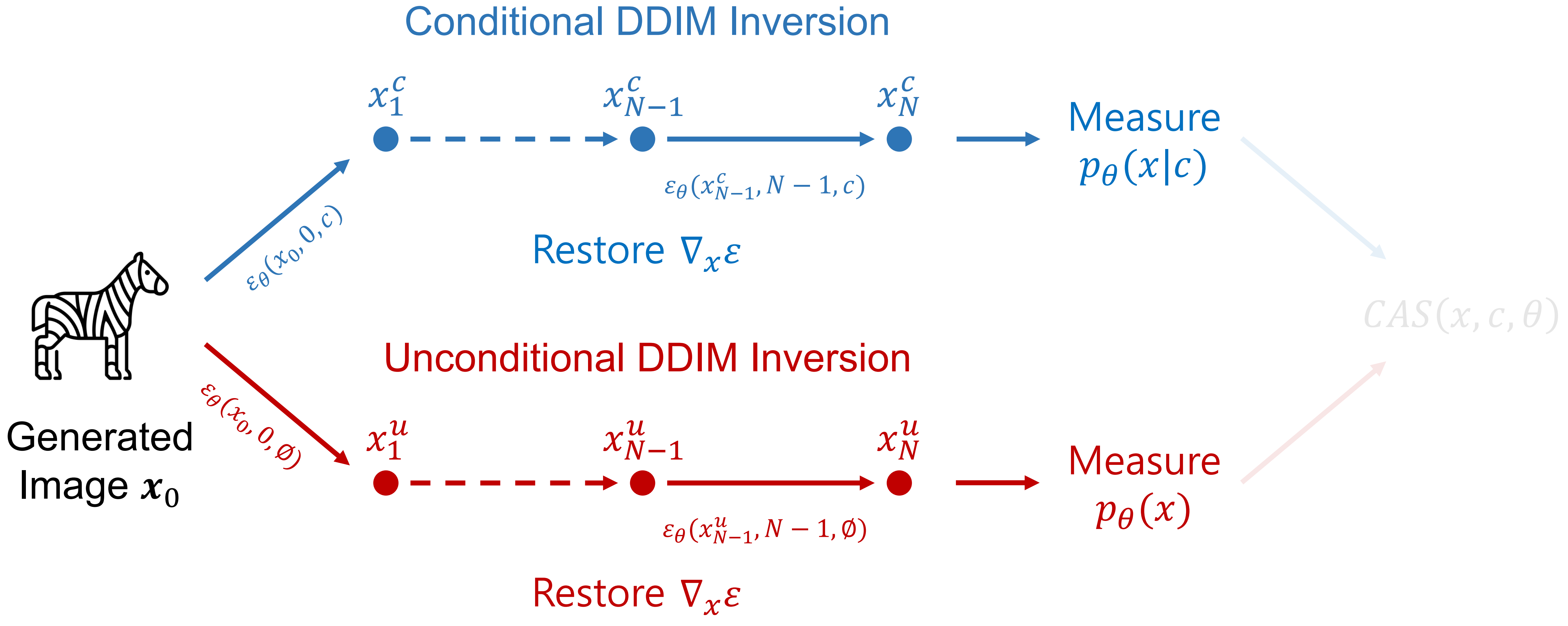
Method in detail:



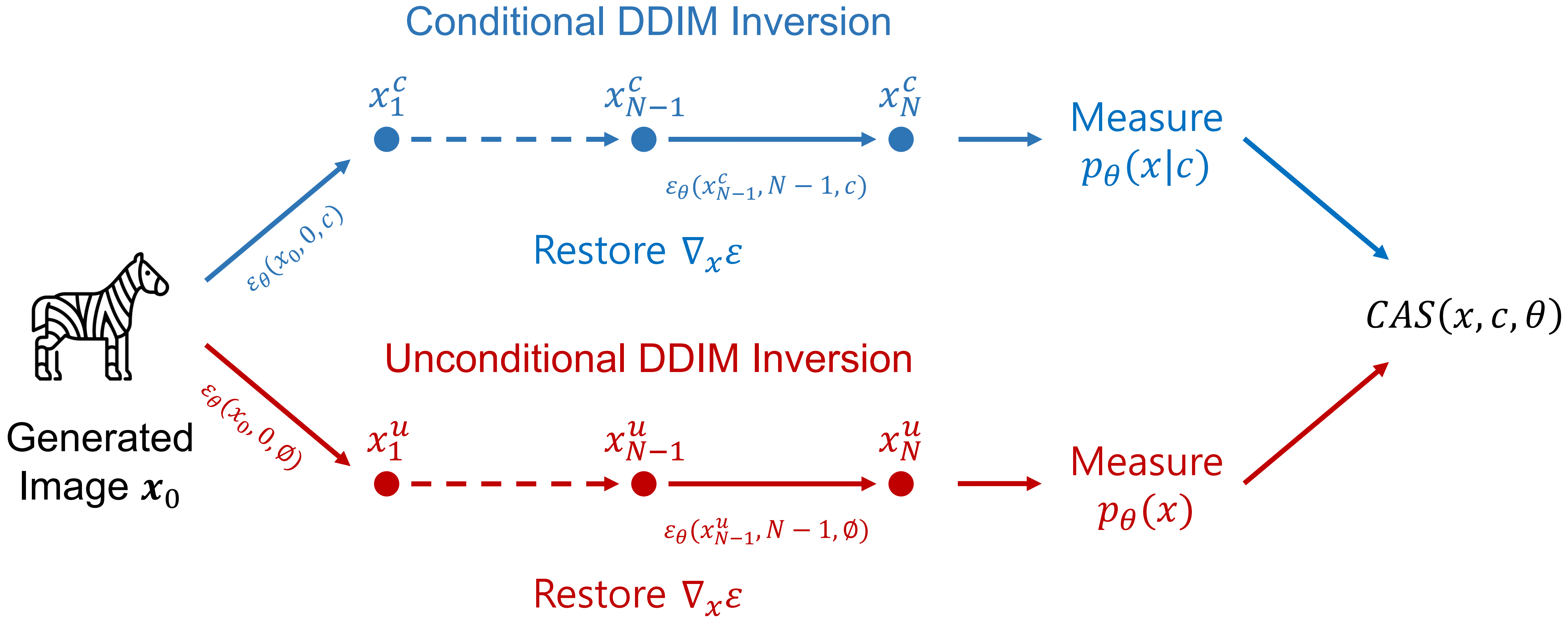
Method in detail:



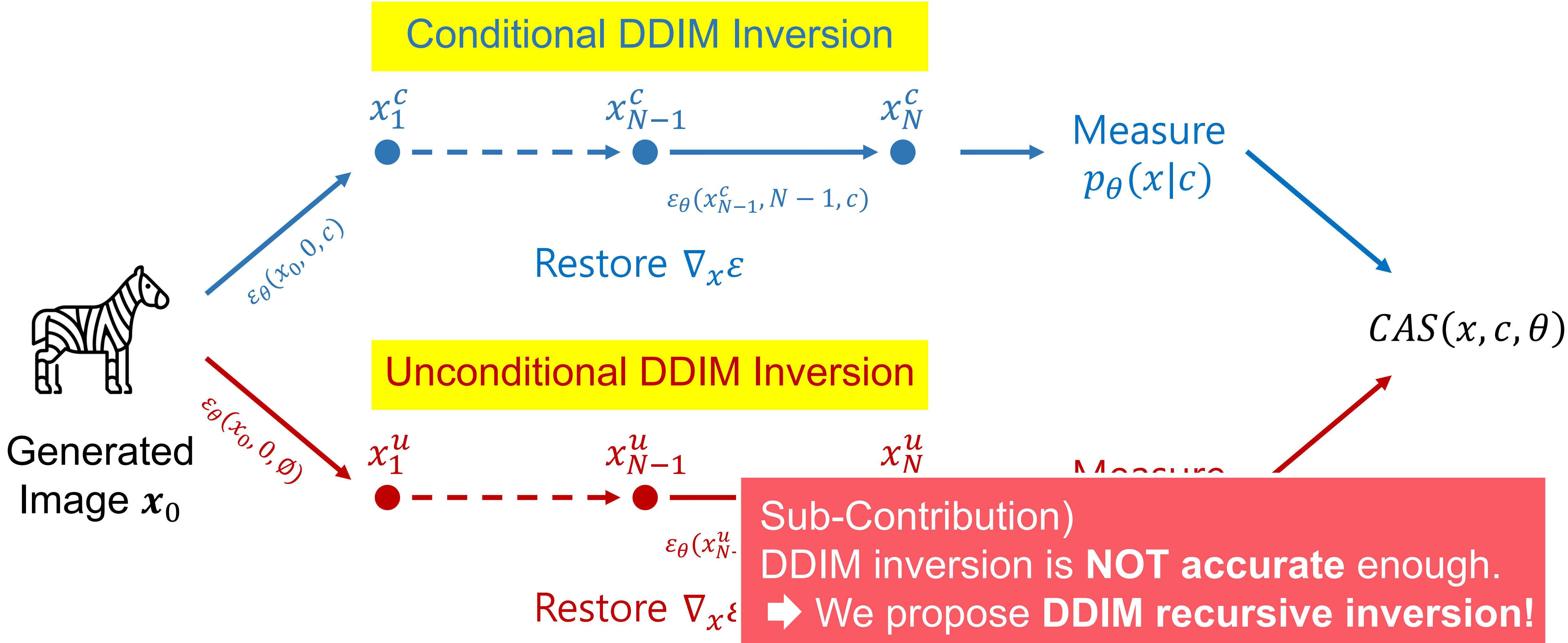
Method in detail:



Method in detail:



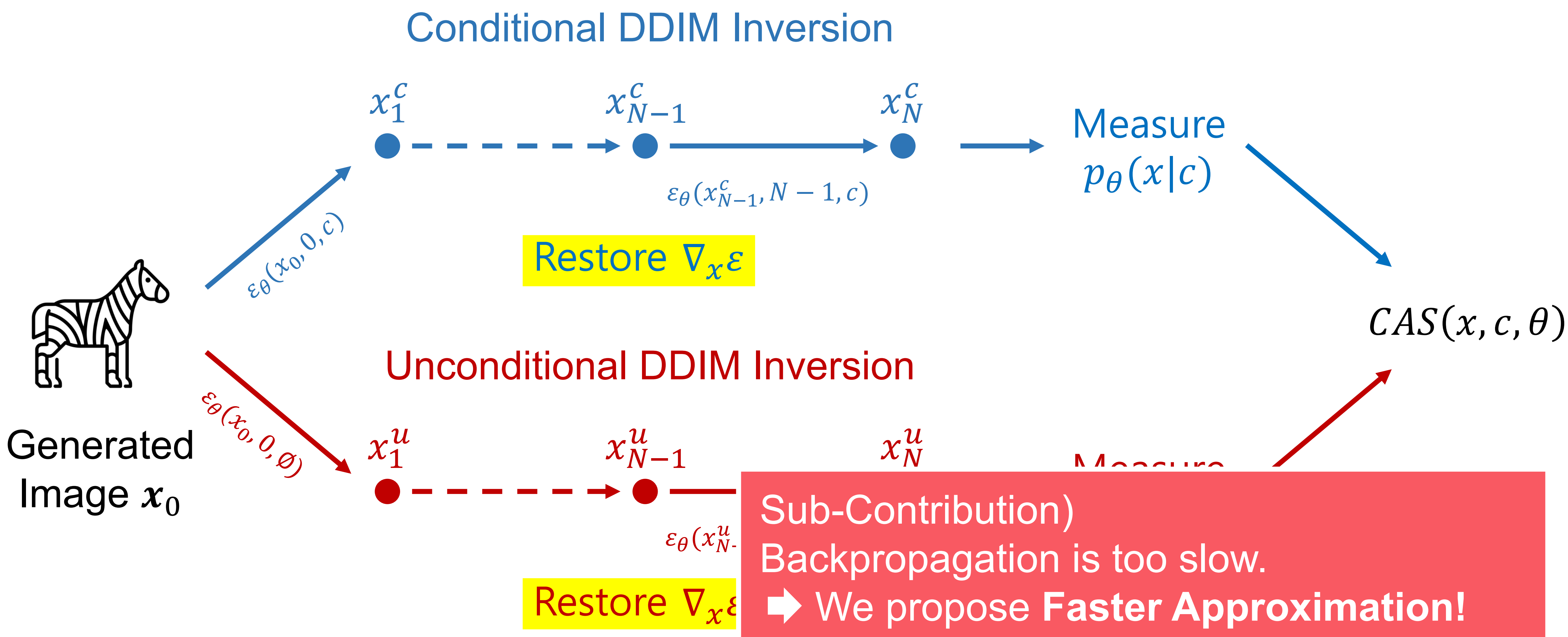
Method in more detail:



Sub-Contribution)
DDIM inversion is **NOT accurate** enough.
➡ We propose **DDIM recursive inversion!**

* For more detail, please refer to the main paper

Method in more detail:



Sub-Contribution)
Backpropagation is too slow.
➡ We propose **Faster Approximation!**

* For more detail, please refer to the main paper

Comparison to previous works:

Number of train data used

	CLIP Score [1]	Image Reward [2]	HPS [3]	Pick Score [4]	Ours
Text #	400M	9K	~99K	~584K	**None**
Image #	400M	4~9/text	~25K	~38K	**None**

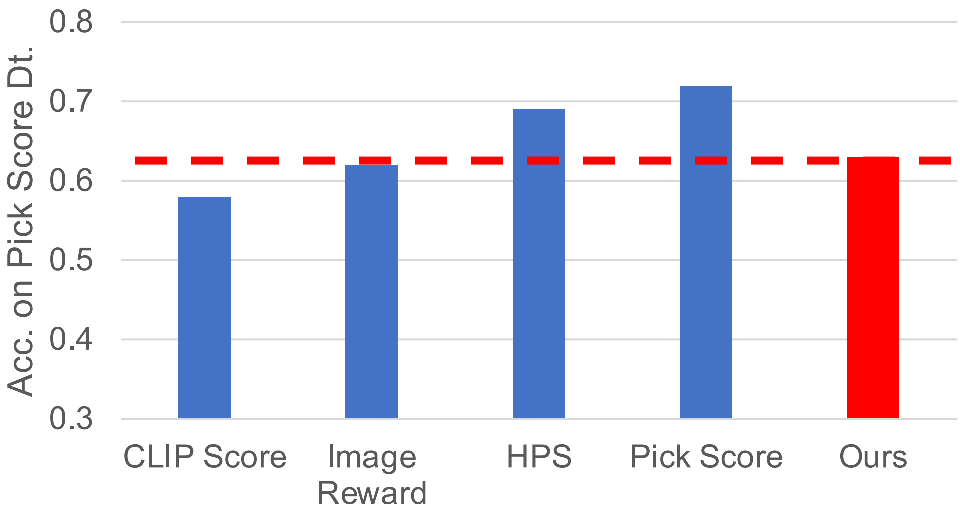
[1] Hessel et. al., CLIPScore: A Reference-free Evaluation Metric for Image Captioning
[2] Su et. al., ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation
[3] Wu et. al., Human Preference Score: Better Aligning Text-to-Image Models with Human Preference
[4] Kirstain et. al., Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation

Comparison to previous works:

Human preference alignment evaluation

Pick Score Dataset

"Western style dog"

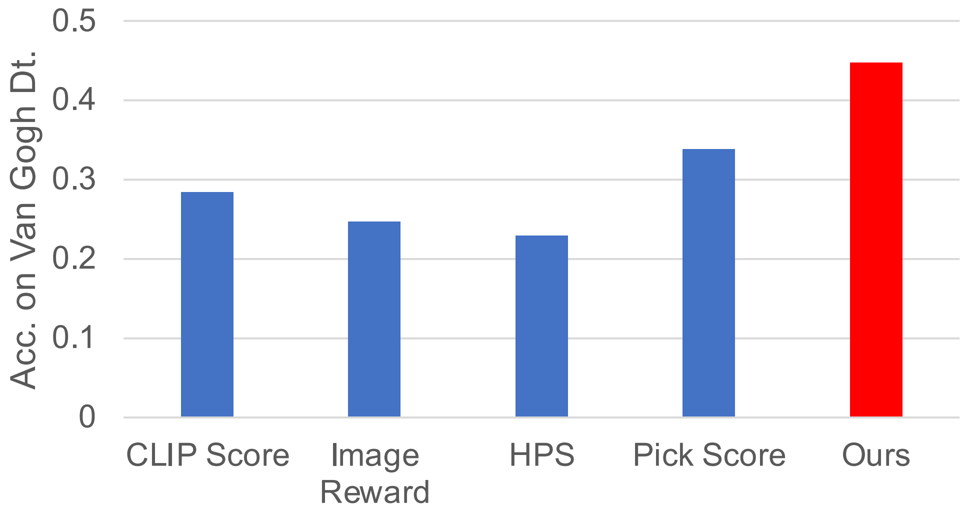


Comparison to previous works:

Human preference alignment evaluation

Van Gogh Dataset

“Van Gogh style, Venezia”



Conclusion

We propose CAS, the universal condition alignment score

To summarize our main contribution, our method

- Leverages conditional probability measured from diffusion model
- Provide DDIM recursive inversion and approximation technique

To summarize our practical benefits, our method

- Is train-free and operates around all domains
- Would be helpful to various domains whose metrics are not defined

To summarize our findings, our method

- Implies that conditional probability space is overfitted to unconditional probability space
- Implies that diffusion models are truly probabilistic

Conclusion

We propose CAS, the universal condition alignment score

To summarize our main contribution, our method

- Leverages conditional probability measured from diffusion model
- Provide DDIM recursive inversion and approximation technique

To summarize our practical benefits, our method

- Is train-free and operates around all domains
- Would be helpful to various domains whose metrics are not defined

To summarize our findings, our method

- Implies that conditional probability space is overfitted to unconditional probability space
- Implies that diffusion models are truly probabilistic

Conclusion

We propose CAS, the universal condition alignment score

To summarize our main contribution, our method

- Leverages conditional probability measured from diffusion model
- Provide DDIM recursive inversion and approximation technique

To summarize our practical benefits, our method

- Is train-free and operates around all domains
- Would be helpful to various domains whose metrics are not defined

To summarize our findings, our method

- Implies that conditional probability space is overfitted to unconditional probability space
- Implies that diffusion models are truly probabilistic

Thank You