NEURAL INFORMATION PROCESSING SYSTEMS

# Diffusion Adaptive Text Embedding for Text-to-Image Diffusion Models

**Byeonghu Na**[1], Minsang Park[1], Gyuwon Sim[1], Donghyeok Shin[1], HeeSun Bae[1], Mina Kang[1], Se Jung Kwon[2], Wanmo Kang[1], Il-Chul Moon[1,3]

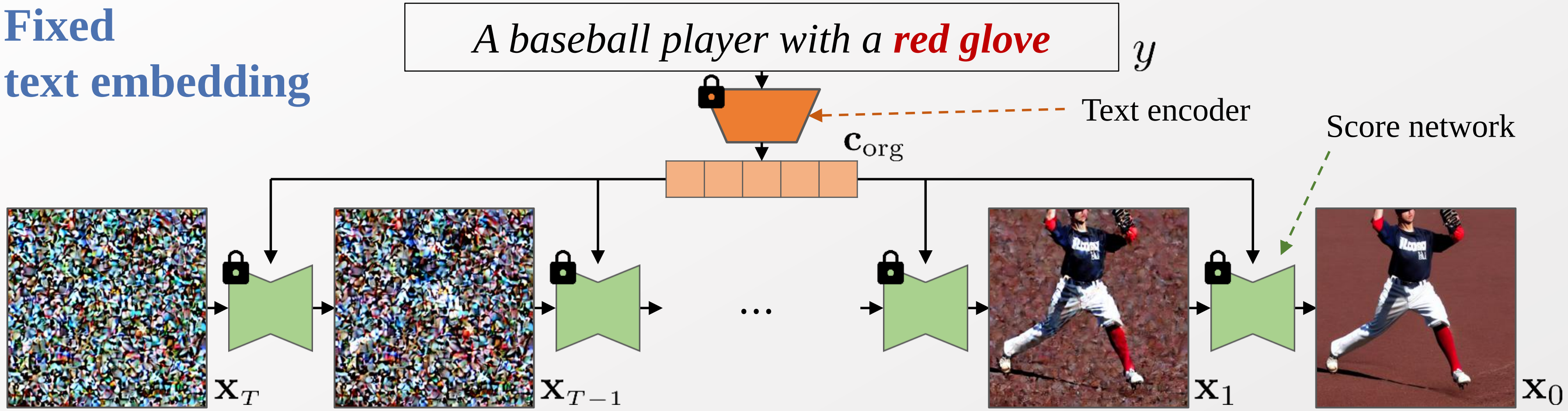
1 KAIST

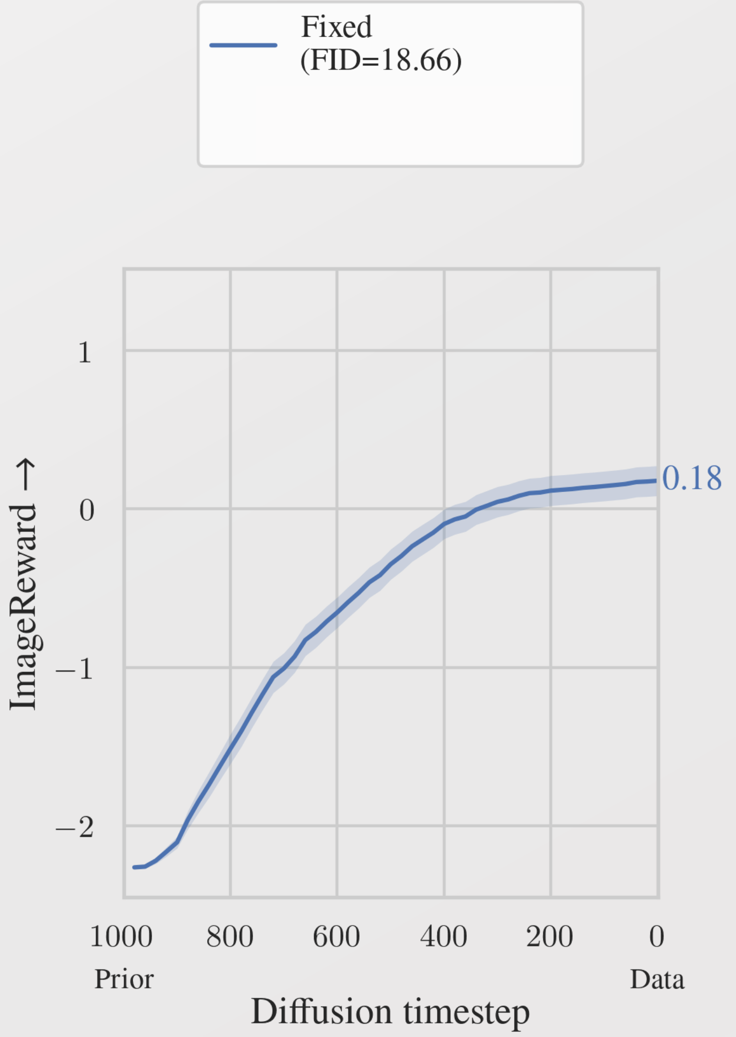
2 NAVER Cloud

3 summary.ai

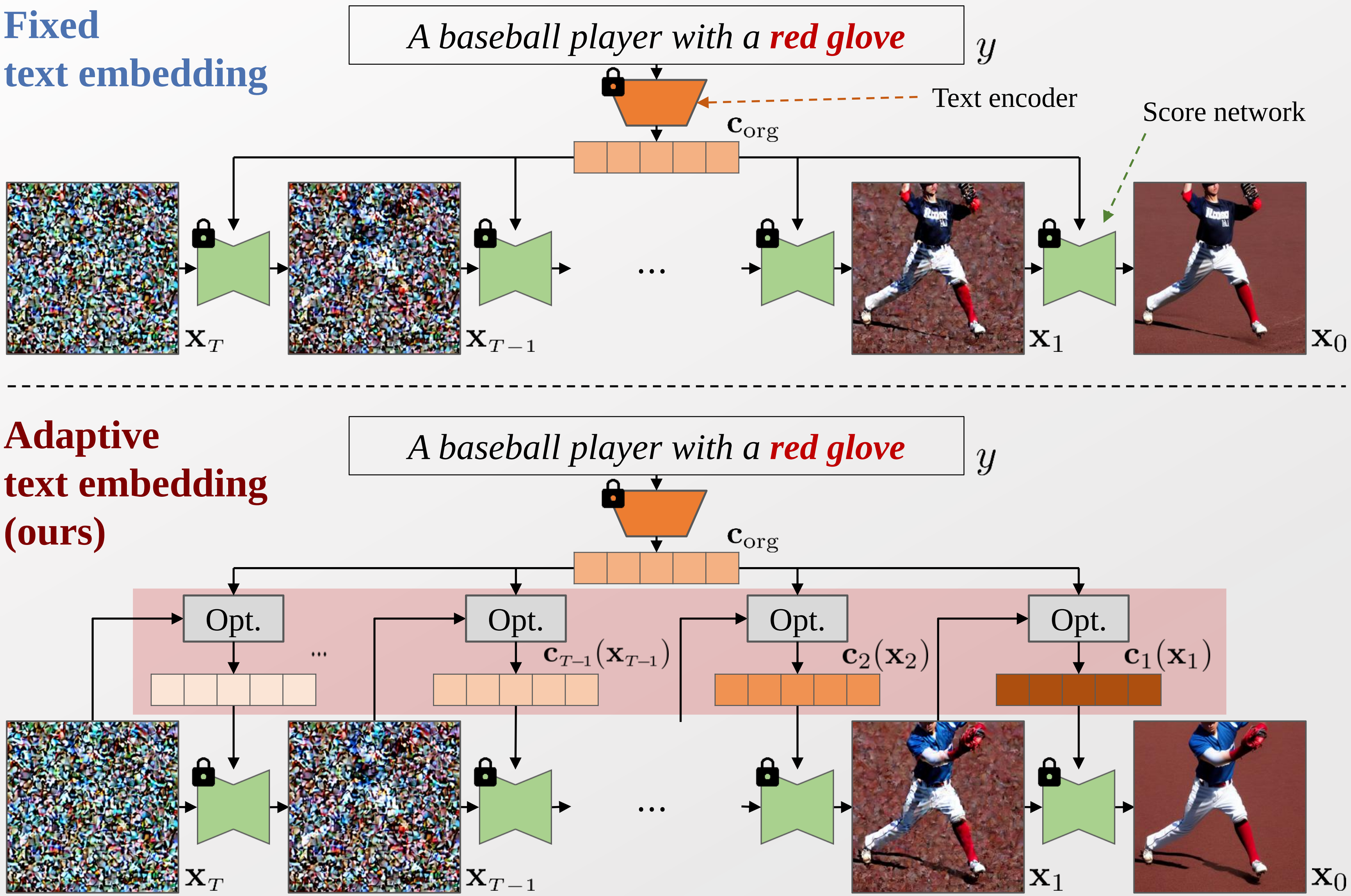
# **Overview** Diffusion Adaptive Text Embedding (DATE)

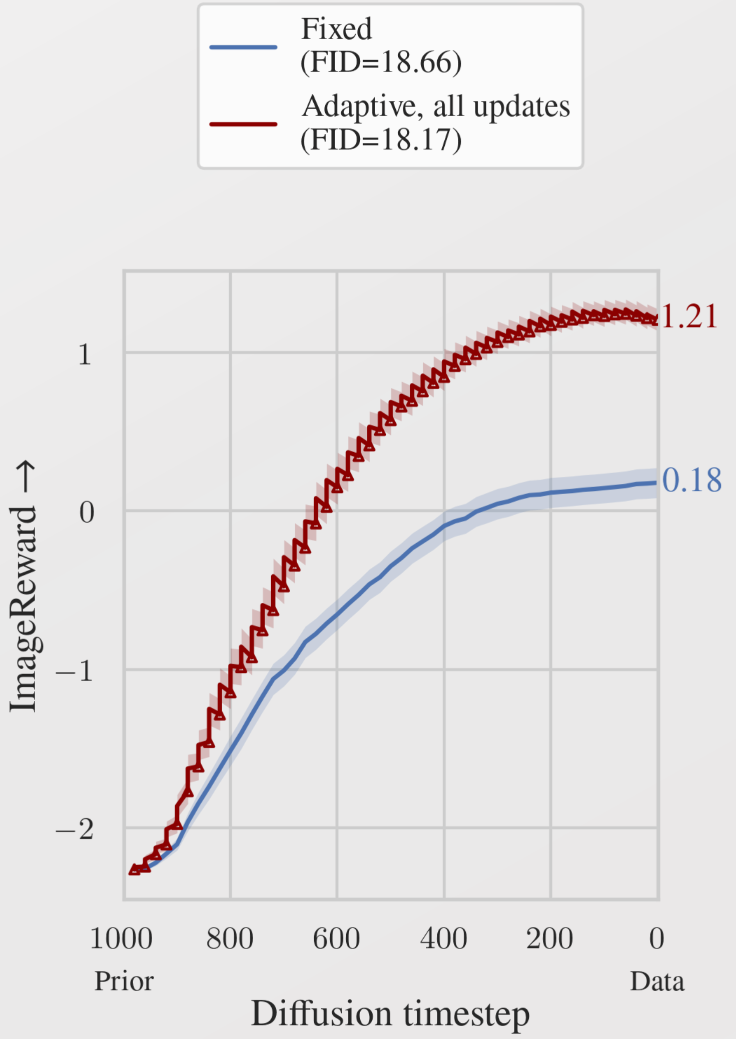
## Fixed text embedding



# **Overview** Diffusion Adaptive Text Embedding (DATE)




KAIST

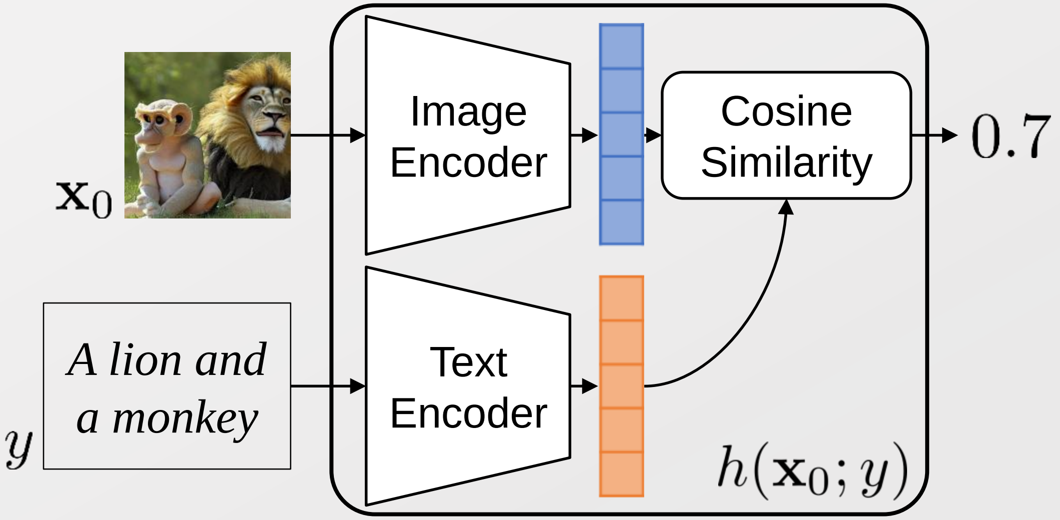
# Text-Conditioned Evaluation Function

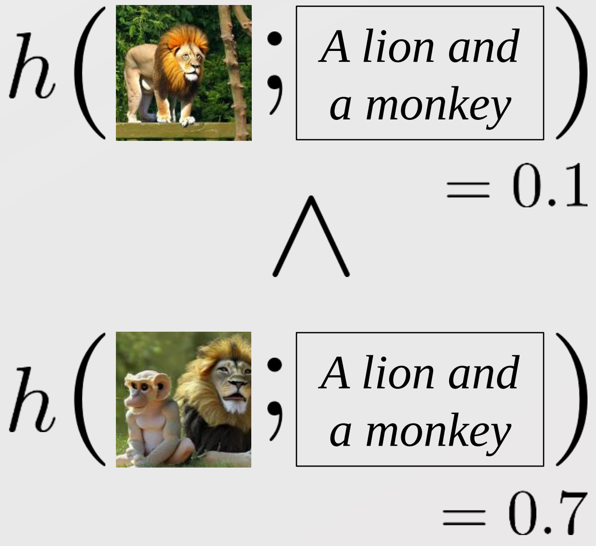
- Text-conditioned evaluation function $h(\mathbf{x}_0; y) \in \mathbb{R}$ 🡐 **Reward function**
  - Evaluate the quality of a generated image in the data space based on a given text condition.
  - Using metrics commonly applied in text-to-image generation evaluation, e.g., CLIP score, ImageReward.

- Limitation of previous usage
  - Only for evaluation purpose without being incorporated into the sampling process.
  - Only for evaluation on the final sample at diffusion timestep 0.

➔ We directly leverage $h$ as the learning objective during the intermediate periods of sampling process.

*Text-conditioned evaluation function (CLIP score)*




- $\mathbf{x}_0$: image
- $y$: text condition

# Diffusion Adaptive Text Embedding (DATE)

- Goal: Find the adaptive text embedding $\boldsymbol{c}_{1:T}$ that maximizes the reward $h$ from the samples generated by the diffusion sampling process $p_\theta$

$$\max_{\mathbf{c}_{1:T}} \mathbb{E}_{\underbrace{\mathbf{x}_T \sim p_T, \mathbf{x}_{0:T-1} \sim \prod_{\tau=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}_{\tau-1} \mid \mathbf{x}_\tau, \mathbf{c}_\tau)}_{\text{Diffusion sampling process}}} [h(\mathbf{x}_0; y)]$$

- $\mathbf{x}_0$: image
- $y$: text condition
- $\mathbf{c}_t$: text embedding at $t$

# Diffusion Adaptive Text Embedding (DATE)

- Goal: Find the adaptive text embedding $\boldsymbol{c}_{1:T}$ that maximizes the reward $h$ from the samples generated by the diffusion sampling process $p_\theta$

$$\max_{\mathbf{c}_{1:T}} \mathbb{E}_{\underbrace{\mathbf{x}_T \sim p_T, \mathbf{x}_{0:T-1} \sim \prod_{\tau=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}_{\tau-1} \mid \mathbf{x}_\tau, \mathbf{c}_\tau)}_{\text{Diffusion sampling process}}} [h(\mathbf{x}_0; y)]$$

*Introduce the constraints for tractable optimization*

*Sequential optimization like diffusion sampling process*

$$\max_{\mathbf{c}_1} \cdots \max_{\mathbf{c}_t} \cdots \max_{\mathbf{c}_T} \mathbb{E}_{\mathbf{x}_T \sim p_T, \mathbf{x}_{0:T-1} \sim \prod_{\tau=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}_{\tau-1} \mid \mathbf{x}_\tau, \mathbf{c}_\tau)} [h(\mathbf{x}_0; y)]$$

*Finding the shared embedding for remaining sampling steps & Close to the original embedding*

$$\max_{\mathbf{c}_1 \in \mathcal{C}_1} \cdots \max_{\mathbf{c}_t \in \mathcal{C}_t} \cdots \max_{\mathbf{c}_T \in \mathcal{C}_T} \mathbb{E}_{\mathbf{x}_T \sim p_T, \mathbf{x}_{0:T-1} \sim \prod_{\tau=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}_{\tau-1} \mid \mathbf{x}_\tau, \mathbf{c}_\tau)} [h(\mathbf{x}_0; y)]$$

- $\mathcal{C}_t := \{\mathbf{c}_t : \|\mathbf{c}_t - \mathbf{c}_{\text{org}}\|_2 \leq \rho, \mathbf{c}_\tau = \mathbf{c}_t \ \forall \tau < t\}$
- $\mathbf{c}_{\text{org}}$: original text embedding
- $\rho$: scale hyperparameter

- $\mathbf{x}_0$: image
- $y$: text condition
- $\mathbf{c}_t$: text embedding at $t$

# Diffusion Adaptive Text Embedding (DATE)

- Goal: Find the adaptive text embedding $\boldsymbol{c}_{1:T}$ that maximizes the reward $h$ from the samples generated by the diffusion sampling process $p_\theta$

$$\max_{\mathbf{c}_{1:T}} \mathbb{E}_{\underbrace{\mathbf{x}_T \sim p_T, \mathbf{x}_{0:T-1} \sim \prod_{\tau=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}_{\tau-1} \mid \mathbf{x}_\tau, \mathbf{c}_\tau)}_{\text{Diffusion sampling process}}} [h(\mathbf{x}_0; y)]$$

*Introduce the constraints for tractable optimization*

From t = $T$ to 1,

$$\max_{\mathbf{c}_t : \|\mathbf{c}_t - \mathbf{c}_{\text{org}}\|_2 \le \rho} \mathbb{E}_{\mathbf{x}_0 \sim p_{\boldsymbol{\theta}}(\mathbf{x}_0 \mid \mathbf{x}_t, \mathbf{c}_t)} [h(\mathbf{x}_0; y)]$$

- $\mathbf{c}_{\text{org}}$: original text embedding
- $\rho$: scale hyperparameter

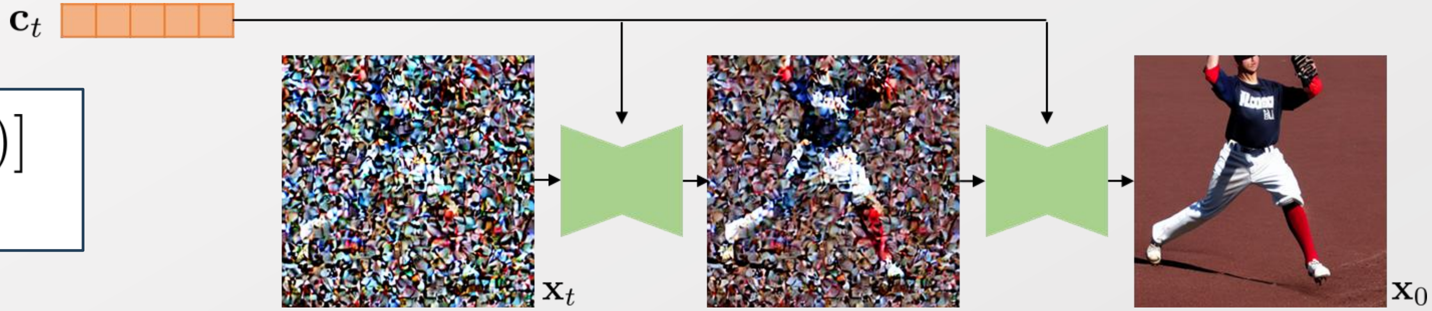
➔ Maximize the expectation of the reward with respect to $p_\theta(x_0|x_t, c_t)$

- $\mathbf{x}_0$: image
- $y$: text condition
- $\mathbf{c}_t$: text embedding at $t$

# Diffusion Adaptive Text Embedding (DATE)

- But, this objective is computationally expensive due to the multiple network evaluations for sampling.

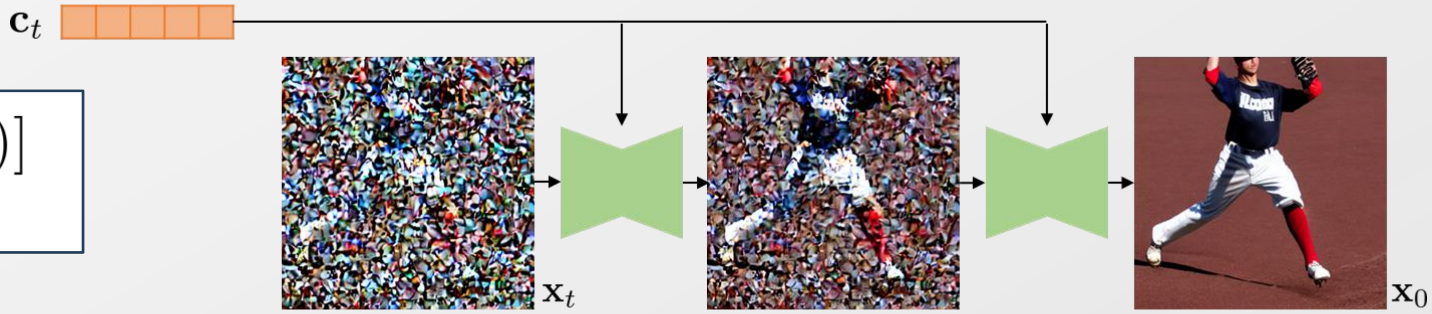
$$\max_{\mathbf{c}_t : \|\mathbf{c}_t - \mathbf{c}_{\text{org}}\| \leq \rho} \mathbb{E}_{\mathbf{x}_0 \sim p_{\boldsymbol{\theta}}(\mathbf{x}_0 \mid \mathbf{x}_t, \mathbf{c}_t)} [h(\mathbf{x}_0; y)]$$



# Diffusion Adaptive Text Embedding (DATE)

- But, this objective is computationally expensive due to the multiple network evaluations for sampling.

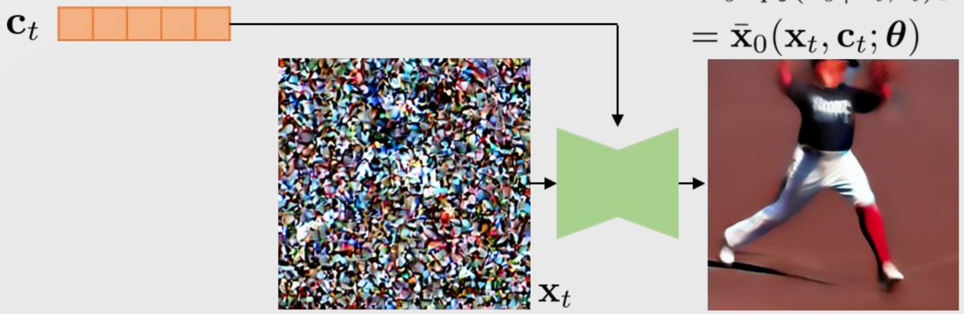
$$\max_{\mathbf{c}_t : \|\mathbf{c}_t - \mathbf{c}_{\text{org}}\| \le \rho} \mathbb{E}_{\mathbf{x}_0 \sim p_{\boldsymbol{\theta}}(\mathbf{x}_0 \mid \mathbf{x}_t, \mathbf{c}_t)} [h(\mathbf{x}_0; y)]$$




*Apply first-order Taylor approx. of $h$ around $\bar{x}_0$*

$$\mathbb{E}_{\mathbf{x}_0 \sim p_{\boldsymbol{\theta}}(\mathbf{x}_0 \mid \mathbf{x}_t, \mathbf{c}_t)} [h(\mathbf{x}_0; y)] \approx h(\underbrace{\mathbb{E}_{\mathbf{x}_0 \sim p_{\boldsymbol{\theta}}(\mathbf{x}_0 \mid \mathbf{x}_t, \mathbf{c}_t)} [\mathbf{x}_0]}_{:= \bar{\mathbf{x}}_0(\mathbf{x}_t, \mathbf{c}_t; \boldsymbol{\theta})}; y)$$

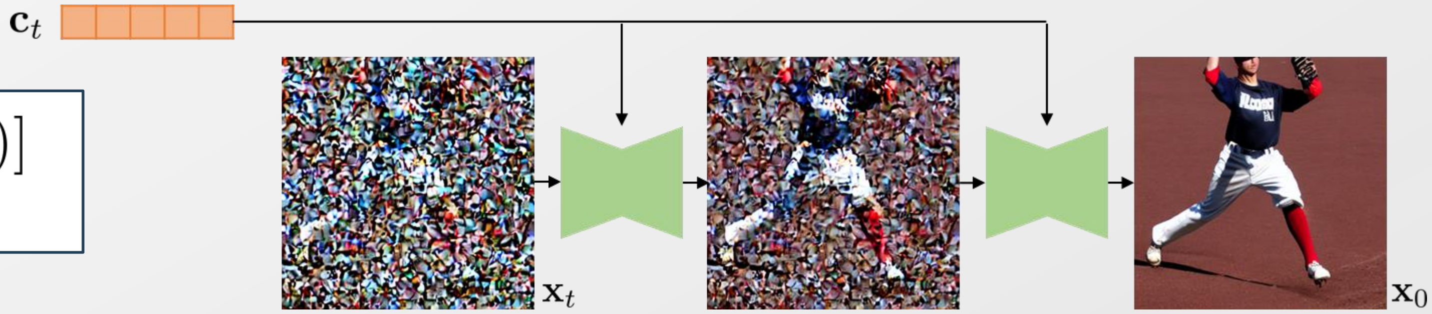
$$\max_{\mathbf{c}_t : \|\mathbf{c}_t - \mathbf{c}_{\text{org}}\| \le \rho} h(\bar{\mathbf{x}}_0(\mathbf{x}_t, \mathbf{c}_t; \boldsymbol{\theta}); y) =: h_t(\mathbf{x}_t, \mathbf{c}_t; y, \boldsymbol{\theta})$$




- Maximize the reward on the mean predicted image $\bar{x}_0$ given current perturbed image $x_t$ and text embedding $c_t$.

# Diffusion Adaptive Text Embedding (DATE)

- But, this objective is computationally expensive due to the multiple network evaluations for sampling.

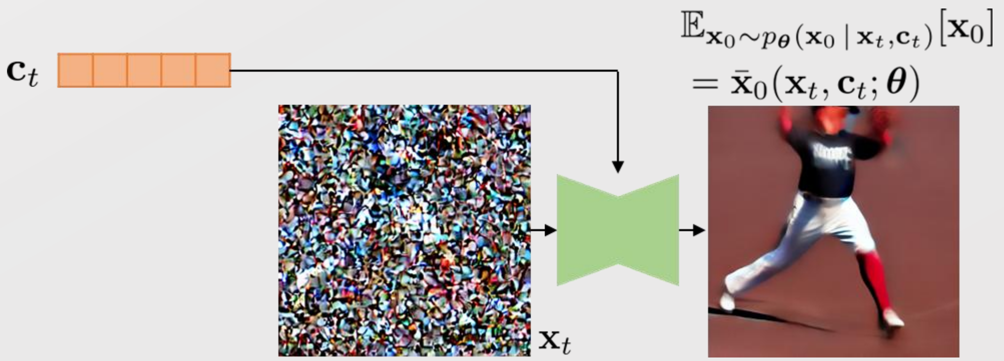



$$\max_{\mathbf{c}_t : \| \mathbf{c}_t - \mathbf{c}_{\text{org}} \| \leq \rho} \mathbb{E}_{\mathbf{x}_0 \sim p_{\boldsymbol{\theta}}(\mathbf{x}_0 \mid \mathbf{x}_t, \mathbf{c}_t)} [h(\mathbf{x}_0; y)]$$

*Apply first-order Taylor approx. of $h$ around $\bar{x}_0$*

$$\mathbb{E}_{\mathbf{x}_0 \sim p_{\boldsymbol{\theta}}(\mathbf{x}_0 \mid \mathbf{x}_t, \mathbf{c}_t)} [h(\mathbf{x}_0; y)] \approx h(\underbrace{\mathbb{E}_{\mathbf{x}_0 \sim p_{\boldsymbol{\theta}}(\mathbf{x}_0 \mid \mathbf{x}_t, \mathbf{c}_t)} [\mathbf{x}_0]}_{:= \bar{\mathbf{x}}_0(\mathbf{x}_t, \mathbf{c}_t; \boldsymbol{\theta})}; y)$$

$$\max_{\mathbf{c}_t : \| \mathbf{c}_t - \mathbf{c}_{\text{org}} \| \leq \rho} h(\bar{\mathbf{x}}_0(\mathbf{x}_t, \mathbf{c}_t; \boldsymbol{\theta}); y) =: h_t(\mathbf{x}_t, \mathbf{c}_t; y, \boldsymbol{\theta})$$




- Maximize the reward on the mean predicted image $\bar{x}_0$ given current perturbed image $x_t$ and text embedding $c_t$.
- Using the Tweedie's formula, the mean predicted image $\bar{x}_0$ can be computed via a single score network evaluation.

$$\bar{\mathbf{x}}_0(\mathbf{x}_t, \mathbf{c}_t; \boldsymbol{\theta}) = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t + (1 - \bar{\alpha}_t)\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{c}_t, t))$$

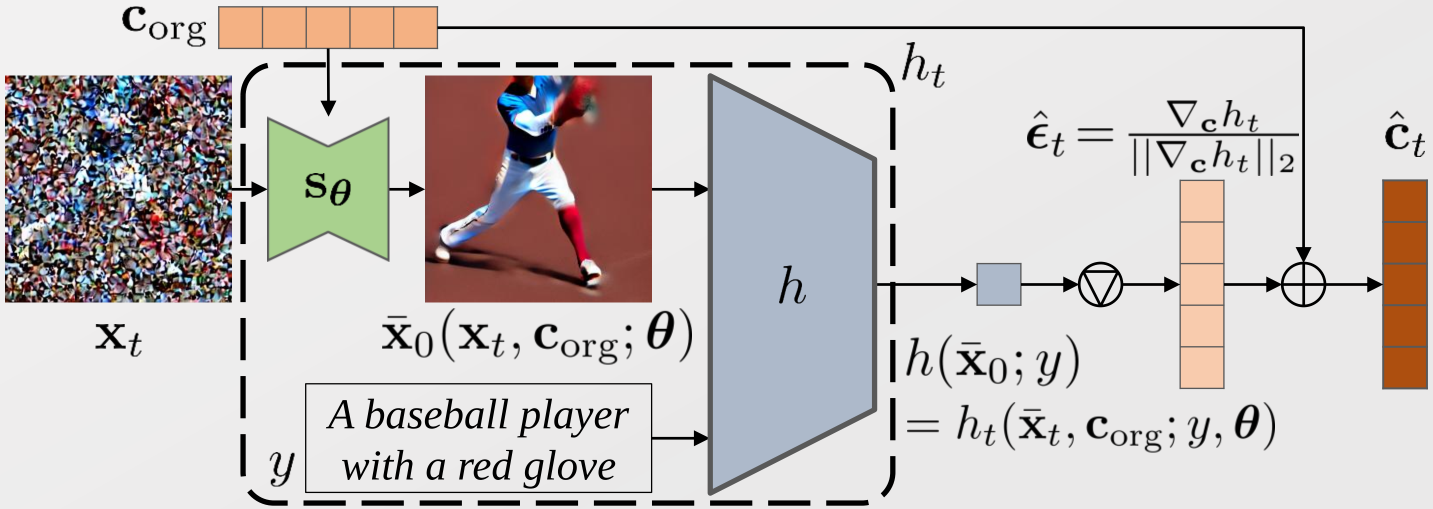
# Update Process of DATE

- To update the text embeddings for each timestep, we use a first-order Taylor approximation for computational efficiency, inspired by the inner maximization of sharpness-aware minimization.

$$\mathbf{c}_t = \mathbf{c}_{\text{org}} + \boldsymbol{\epsilon}_t$$

$$\begin{aligned}\boldsymbol{\epsilon}_t^* &:= \underset{||\boldsymbol{\epsilon}_t||_2 \leq \rho}{\arg\max}\, h_t(\mathbf{x}_t, \mathbf{c}_{\text{org}} + \boldsymbol{\epsilon}_t; y, \boldsymbol{\theta}) \\ &\approx \underset{||\boldsymbol{\epsilon}_t||_2 \leq \rho}{\arg\max} \left\{ h_t(\mathbf{x}_t, \mathbf{c}_{\text{org}}; y, \boldsymbol{\theta}) + \boldsymbol{\epsilon}_t^{\mathrm{T}} \nabla_{\mathbf{c}} h_t(\mathbf{x}_t, \mathbf{c}_{\text{org}}; y, \boldsymbol{\theta}) \right\} \\ &= \underset{||\boldsymbol{\epsilon}_t||_2 \leq \rho}{\arg\max}\, \boldsymbol{\epsilon}_t^{\mathrm{T}} \nabla_{\mathbf{c}} h_t(\mathbf{x}_t, \mathbf{c}_{\text{org}}; y, \boldsymbol{\theta}) =: \hat{\boldsymbol{\epsilon}}_t\end{aligned}$$

$$\hat{\mathbf{c}}_t = \mathbf{c}_{\text{org}} + \rho \frac{\nabla_{\mathbf{c}} h_t(\mathbf{x}_t, \mathbf{c}_{\text{org}}; y, \boldsymbol{\theta})}{||\nabla_{\mathbf{c}} h_t(\mathbf{x}_t, \mathbf{c}_{\text{org}}; y, \boldsymbol{\theta})||_2}$$




- ▽: the normalized gradient with respect to *c*
- ⊕: summation

# **Theoretical Analysis:** Performance Guarantee

- Both unconstrained and constrained optimizations of the text embedding produce a better text embedding than the fixed text embedding.

$$\max_{\mathbf{c}_{1:T}} \mathbb{E}_{\mathbf{x}_T \sim p_T, \mathbf{x}_{0:T-1} \sim \prod_{\tau=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}_{\tau-1} \mid \mathbf{x}_\tau, \mathbf{c}_\tau)} [h(\mathbf{x}_0; y)]$$ // Unconstrained optimization

$$= \max_{\mathbf{c}_1} \cdots \max_{\mathbf{c}_t} \cdots \max_{\mathbf{c}_T} \mathbb{E}_{\mathbf{x}_T \sim p_T, \mathbf{x}_{0:T-1} \sim \prod_{\tau=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}_{\tau-1} \mid \mathbf{x}_\tau, \mathbf{c}_\tau)} [h(\mathbf{x}_0; y)]$$

$$\geq \max_{\mathbf{c}_1 \in \mathcal{C}_1} \cdots \max_{\mathbf{c}_t \in \mathcal{C}_t} \cdots \max_{\mathbf{c}_T \in \mathcal{C}_T} \mathbb{E}_{\mathbf{x}_T \sim p_T, \mathbf{x}_{0:T-1} \sim \prod_{\tau=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}_{\tau-1} \mid \mathbf{x}_\tau, \mathbf{c}_\tau)} [h(\mathbf{x}_0; y)]$$ // Constrained optimization

$$\geq h(\mathbf{c}_{\mathrm{org}}, \cdots, \mathbf{c}_{\mathrm{org}})$$ // Fixed text embedding

- Since DATE is derived by approximating the constrained optimization,
  it is expected to improve the quality of the generated images compared to the fixed embedding.

KAIST

# **Theoretical Analysis:** Influence on Data Space

- How the DATE update influences the perturbed data?

The score function for the text embedding $c_t$ updated by DATE can be expressed as:

$$\nabla_{\mathbf{x}_t} \log p_{\boldsymbol{\theta}}(\mathbf{x}_t \,|\hat{\mathbf{c}}_t) = \underbrace{\nabla_{\mathbf{x}_t} \log p_{\boldsymbol{\theta}}(\mathbf{x}_t \mid \mathbf{c}_{\text{org}})}_{\text{Original score function}} + \rho \nabla_{\mathbf{x}_t} \left\{ \underbrace{\frac{\nabla_{\mathbf{c}} h_t(\mathbf{x}_t, \mathbf{c}_{\text{org}})^T}{||\nabla_{\mathbf{c}} h_t(\mathbf{x}_t, \mathbf{c}_{\text{org}})||_2}}_{\text{Reward}} \underbrace{\nabla_{\mathbf{c}} \log p_{\boldsymbol{\theta}}(\mathbf{x}_t \mid \mathbf{c}_{\text{org}})}_{\text{Model likelihood}} \right\} + O(\rho^2)$$

Alignment

- This guidance improves the alignment between the evaluation function $h_t$ and the model likelihood from the perspective of the text embedding.

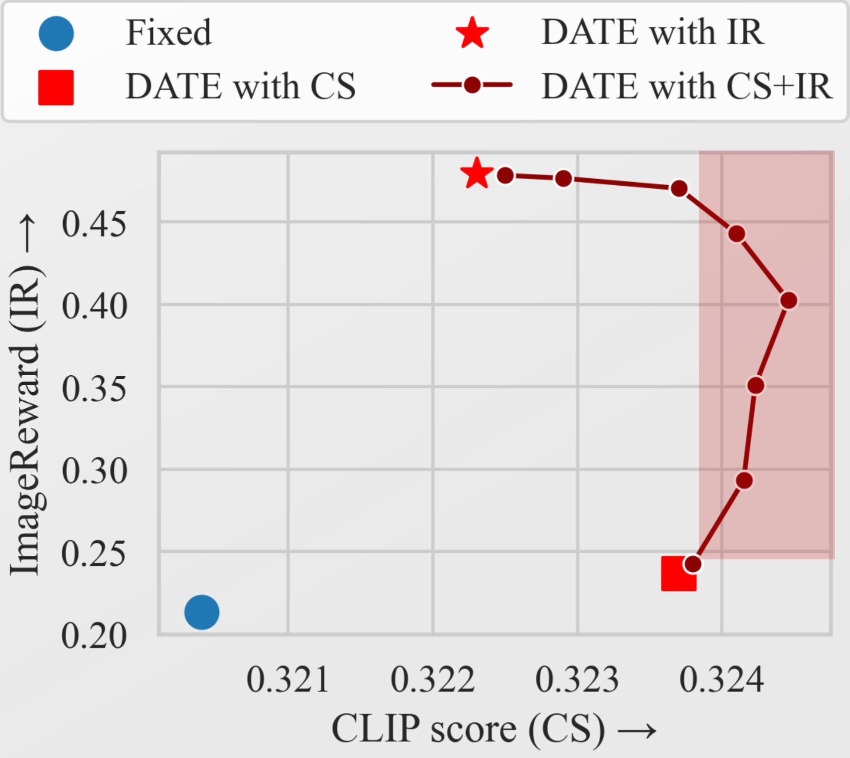
- Embedding-based guidance balances semantic alignment with the underlying model distribution, enhancing prompt fidelity without reducing generation quality.

# Experiments Quantitative Results

- FID: average distance between generated and reference datasets
- CLIP score: text-image embedding alignment score via CLIP
- ImageReward: text-to-image human preference score

| Backbone | Method | Time | FID↓ | CLIP score↑ | ImageReward↑ |
|---|---|---|---|---|---|
| SD v1.5 w/ DDIM | Fixed text embedding (50 steps) | 5.64 | 18.66 | 0.3204 | 0.2132 |
| | Fixed text embedding (70 steps) | 7.87 | 18.27 | 0.3199 | 0.2137 |
| | EBCA | 8.10 | 25.85 | 0.2877 | -0.3128 |
| | Universal Guidance | 8.25 | 18.56 | 0.3216 | 0.2221 |
| | **DATE (50 steps)** | | | | |
| | 10% update with CLIP score | 7.82 | 17.90 | 0.3237 | 0.2364 |
| | all updates with CLIP score | 24.20 | **17.22** | **0.3292** | 0.2277 |
| | 10% update with ImageReward | 7.82 | 18.61 | 0.3224 | 0.4792 |
| | all updates with ImageReward | 24.20 | 18.17 | 0.3224 | **1.2972** |
| PixArt-$\alpha$ w/ DPM-Solver | Fixed text embedding (20 steps) | 4.35 | 31.07 | 0.3201 | 0.8140 |
| | Fixed text embedding (45 steps) | 9.03 | 30.62 | 0.3199 | 0.8174 |
| | **DATE (20 steps)** | | | | |
| | 50% update with CLIP score | 8.93 | **30.55** | **0.3237** | 0.8287 |
| | 50% update with ImageReward | 8.95 | 31.07 | 0.3221 | **0.9514** |

Performance on COCO validation set




Performance of combined metrics

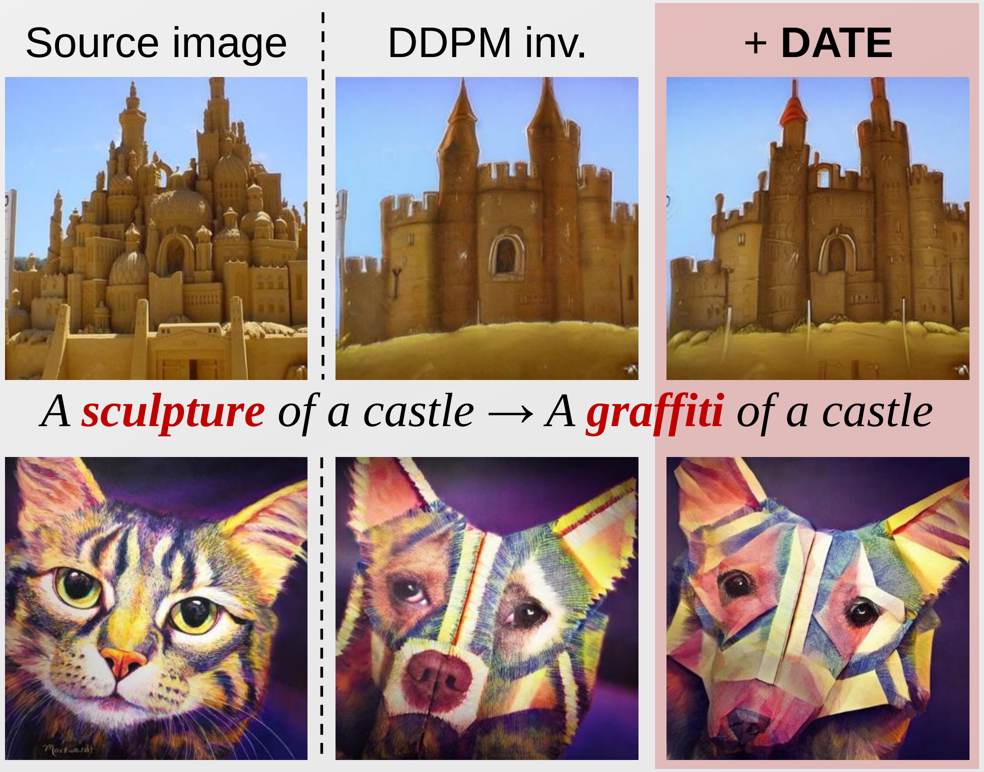
DATE **consistently improves** the semantic alignment and overall image quality.

## Multi-concept generation


*A **lion** and a **monkey***

*A **dog** and a **blue balloon***

## Text-guided image editing




*A **sculpture** of a castle → A **graffiti** of a castle*

*A **cartoon** of a **cat** → An **origami** of a **dog***

# Thank you!

Paper

Code

Contact: byeonghu.na@kaist.ac.kr