High Fidelity Guided Image Synthesis using Latent Diffusion Models

Jaskirat Singh†

Stephen Gould†*

Liang Zheng†*

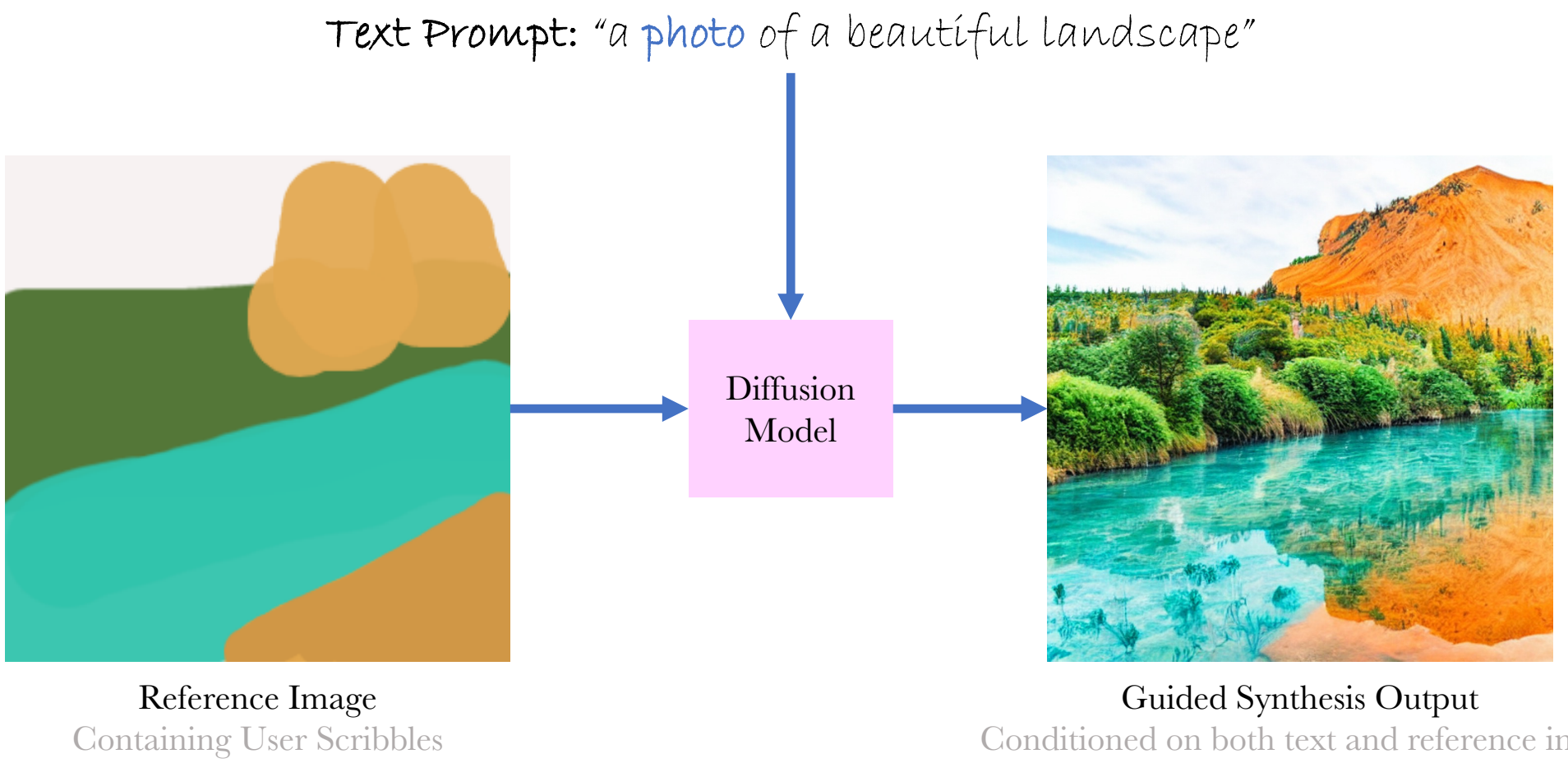
† Australian National University

* Australian Centre for Robotic Vision

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Poster Tag: TUE-PM-179

Guided Image Synthesis with User Scribbles



Text Prompt: "a photo of a beautiful landscape"




Text-Conditioned Image Outputs

Text Prompt: "a photo of a beautiful landscape"




Text-Conditioned Image Outputs

Text Prompt:
"a photo of a beautiful landscape"

Reference Painting




SDEdit

Text Prompt: "a photo of a beautiful landscape"




Text-Conditioned Image Outputs

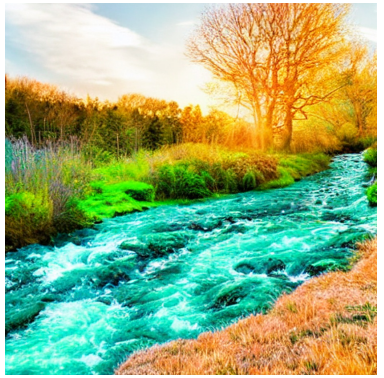
Text Prompt:
"a photo of a beautiful landscape"

Reference Painting




SDEdit






GradOP (Ours)

Controlling Semantics of Different Painting Regions

Text Prompt: "a photo of a beautiful landscape"

GradOP: Constrained Optimization Solution

Text Prompt: "a photo of a red tiger in a green field"

Reference: y

GradOP: Constrained Optimization Solution

Text Prompt: "a photo of a red tiger in a green field"

Reference: y

Target Subspace

conditioned only on text

$\mathcal{S}_{\tau_{text}}$

GradOP: Constrained Optimization Solution

Text Prompt: "a photo of a red tiger in a green field"

Reference: y

$$x^\star = \text{argmin}_x \ \mathcal{L}\left(f(x), y\right)$$
$$s.t. \quad x \in \mathcal{S}_{\tau_{text}}$$

Constrained Optimization Formulation

Target Subspace

conditioned only on text

$\mathcal{S}_{\tau_{text}}$

GradOP: Constrained Optimization Solution

Text Prompt: "a photo of a red tiger in a green field"

Reference: y

$$x^\star = \text{argmin}_x \ \mathcal{L}\left(f(x), y\right)$$
$$s.t. \quad x \in \mathcal{S}_{\tau_{text}}$$

Constrained Optimization Formulation

Target Subspace

conditioned only on text

$\mathcal{S}_{\tau_{text}}$

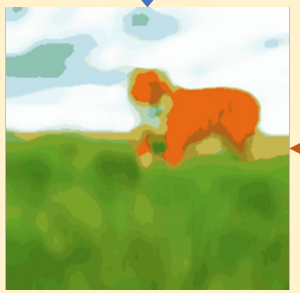
GradOP: Constrained Optimization Solution

Text Prompt: "a photo of a red tiger in a green field"

Target Subspace
conditioned only on text

Reference: y

$\mathcal{L}(f(x), y)$

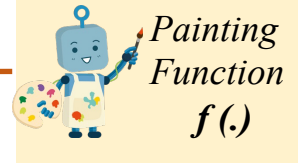


$f(x)$

Painting Reconstruction

$$x^\star = \text{argmin}_x \ \mathcal{L}(f(x), y)$$
$$s.t. \quad x \in \mathcal{S}_{\tau_{text}}$$

Constrained Optimization Formulation



Painting Function $f(.)$

$\mathcal{S}_{\tau_{text}}$

GradOP: Constrained Optimization Solution

Text Prompt: "a photo of a red tiger in a green field"

Reference: y

GradOP Optimization

$\tilde{x}^\star$

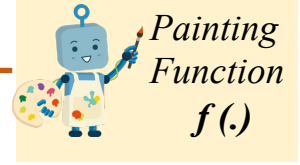
$$\tilde{x}^\star = \text{argmin}_x \; \mathcal{L}(f(x), y) + \gamma \; d(x, x_\tau)$$

Unconstrained Optimization Approximation

$\mathcal{L}(f(x), y)$

$f(x)$

Painting Reconstruction



Target Subspace

conditioned only on text

$\mathcal{S}_{\tau_{text}}$

GradOP: Constrained Optimization Solution

Text Prompt: "a photo of a red tiger in a green field"

Reference: y

GradOP Optimization

$\tilde{x}^\star$

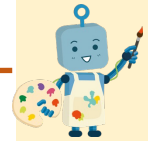
$$\tilde{x}^\star = \text{argmin}_x \; \mathcal{L}(f(x), y) + \gamma \; d(x, x_\tau)$$

Unconstrained Optimization Approximation

$\mathcal{L}(f(x), y)$

$f(x)$

Painting Reconstruction



Painting Function f(.)

Target Subspace

conditioned only on text

$\mathcal{S}_{\tau_{text}}$

GradOP: Constrained Optimization Solution

Text Prompt: "a photo of a red tiger in a green field"

Reference: y

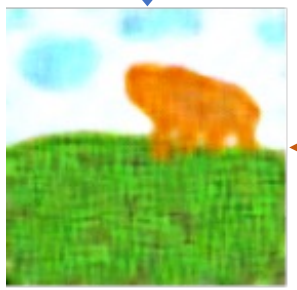
GradOP Optimization

$\tilde{x}^\star$

$$\tilde{x}^\star = \text{argmin}_x \; \mathcal{L}(f(x), y) + \gamma \; d(x, x_\tau)$$

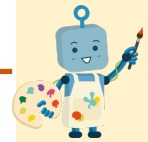
Unconstrained Optimization Approximation

$\mathcal{L}(f(x), y)$



$f(x)$

Painting Reconstruction



Painting Function f(.)

Target Subspace

conditioned only on text

$\mathcal{S}_{\tau_{text}}$

GradOP: Constrained Optimization Solution

Text Prompt: "a photo of a red tiger in a green field"

Reference: y

GradOP Optimization

$\tilde{x}^\star$

$$\tilde{x}^\star = \text{argmin}_x \; \mathcal{L}(f(x), y) + \gamma \; d(x, x_\tau)$$

Unconstrained Optimization Approximation

DDIM Inversion

Target Subspace

conditioned only on text

$\mathcal{S}_{\tau_{text}}$

$\tilde{x}^\star$

Forward Diffusion

x_{t_0}

x_t

$\cdots$

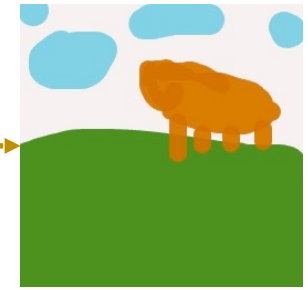
$x^\star$

Output Image

GradOP: Constrained Optimization Solution

Text Prompt: "a photo of a red tiger in a green field"

Reference: y



Target Subspace

conditioned only on text

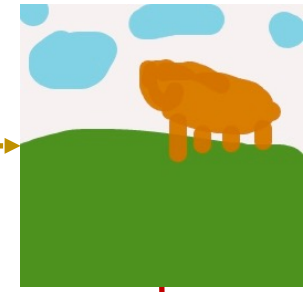
$\mathcal{S}_{\tau_{text}}$

High Domain Gap

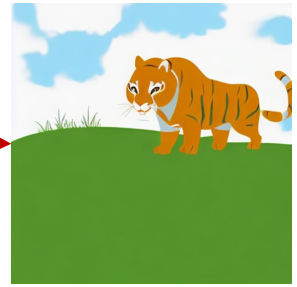
GradOP: Constrained Optimization Solution

Text Prompt: "a photo of a red tiger in a green field"

Reference: y



SDEdit



Target Subspace

conditioned only on text

$\mathcal{S}_{\tau_{text}}$

High Domain Gap

GradOP: Constrained Optimization Solution

Text Prompt: "a photo of a red tiger in a green field"

Reference: y

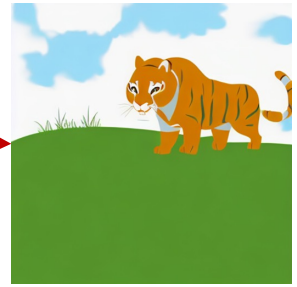
GradOP Optimization

$\tilde{x}^\star$

$$\tilde{x}^\star = \text{argmin}_x \; \mathcal{L}(f(x), y) + \gamma \; d(x, x_\tau)$$

Unconstrained Optimization Approximation

SDEdit



Target Subspace

conditioned only on text

$\mathcal{S}_{\tau_{text}}$

High Domain Gap

GradOP: Constrained Optimization Solution

Text Prompt: "a photo of a red tiger in a green field"

Reference: y

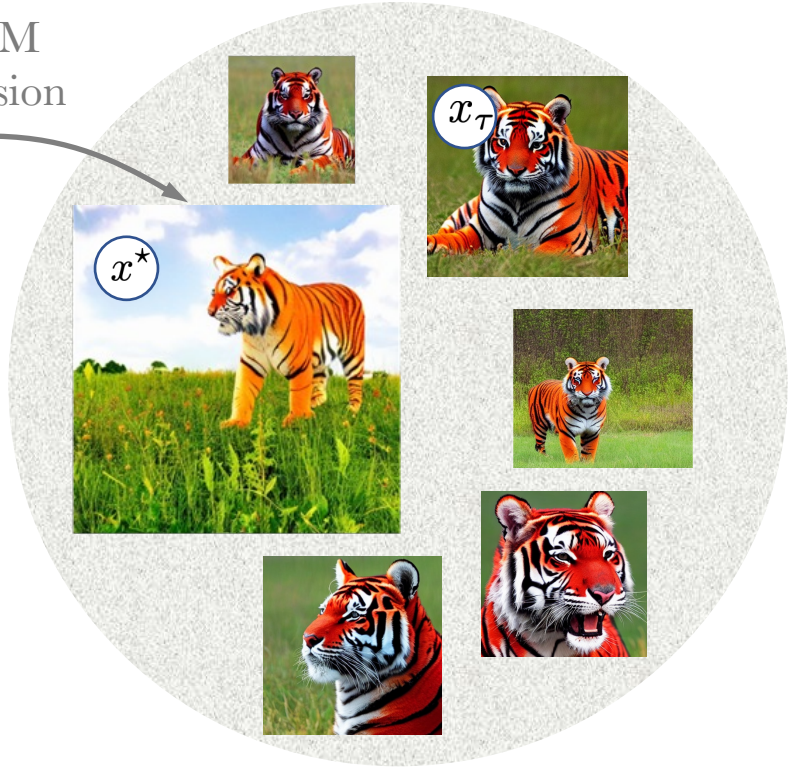
GradOP Optimization

$\tilde{x}^\star$

DDIM Inversion

Target Subspace

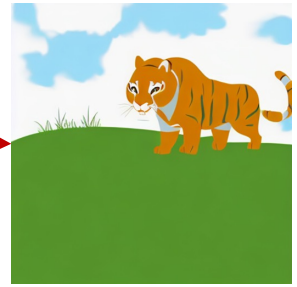
conditioned only on text



$$\tilde{x}^\star = \text{argmin}_x \; \mathcal{L}(f(x), y) + \gamma \; d(x, x_\tau)$$

Unconstrained Optimization Approximation

SDEdit



$\mathcal{S}_{\tau_{text}}$

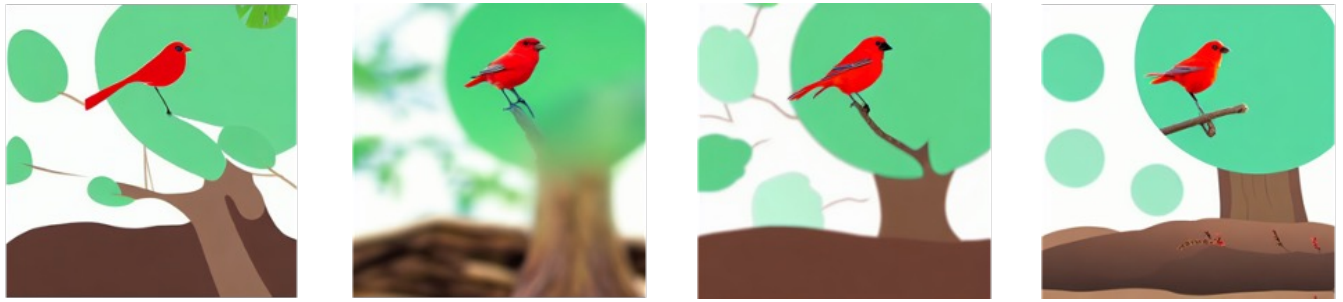
High Domain Gap

Text Prompt: "a photo of a red bird in a tree"

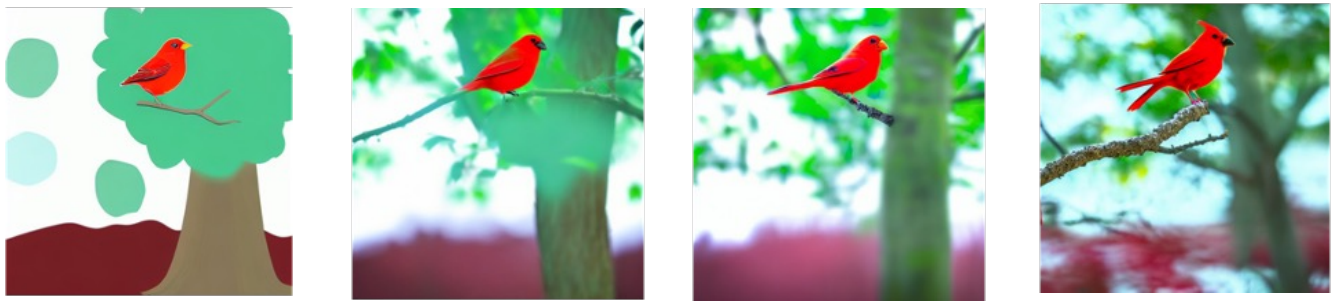
Target Subspace

Text Prompt:
"a photo of a red bird in a tree"

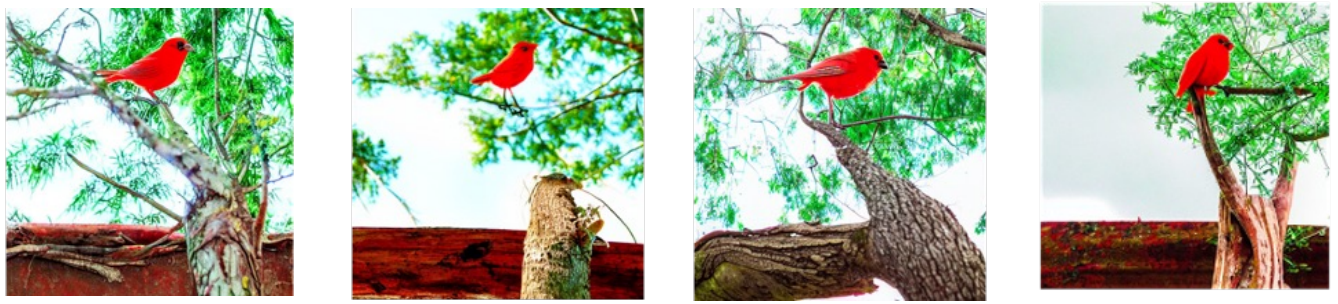
Reference Painting



SDEdit



Iterative Loopback



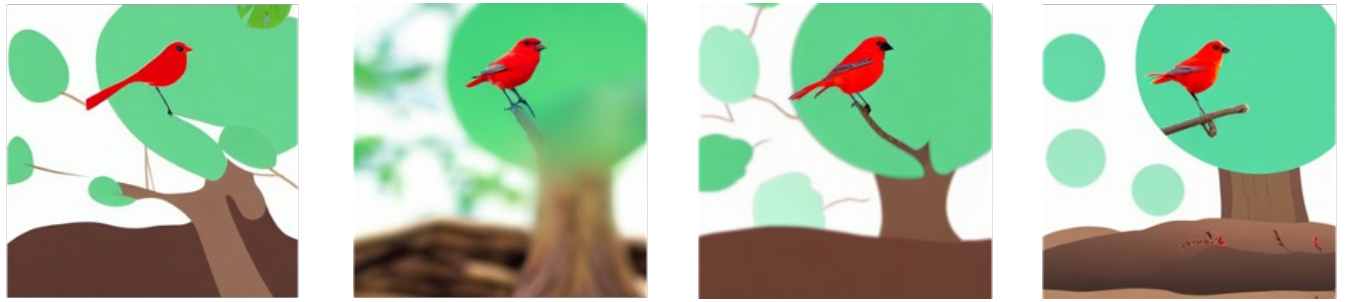
GradOP (Ours)

Text Prompt: "a photo of a red bird in a tree"

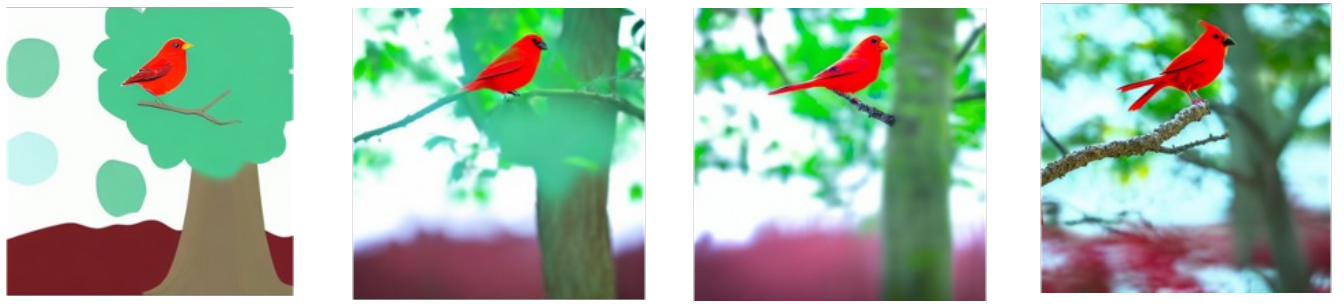
Target Subspace

Text Prompt:
"a photo of a red bird in a tree"

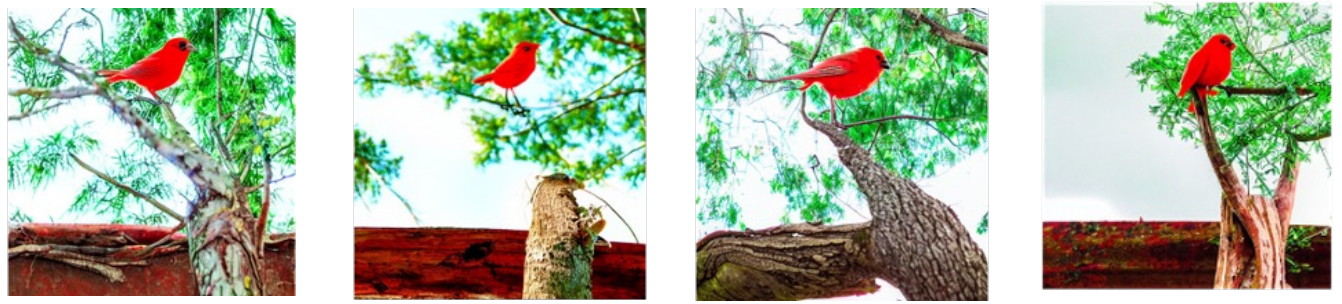
Reference Painting



SDEdit



Iterative Loopback



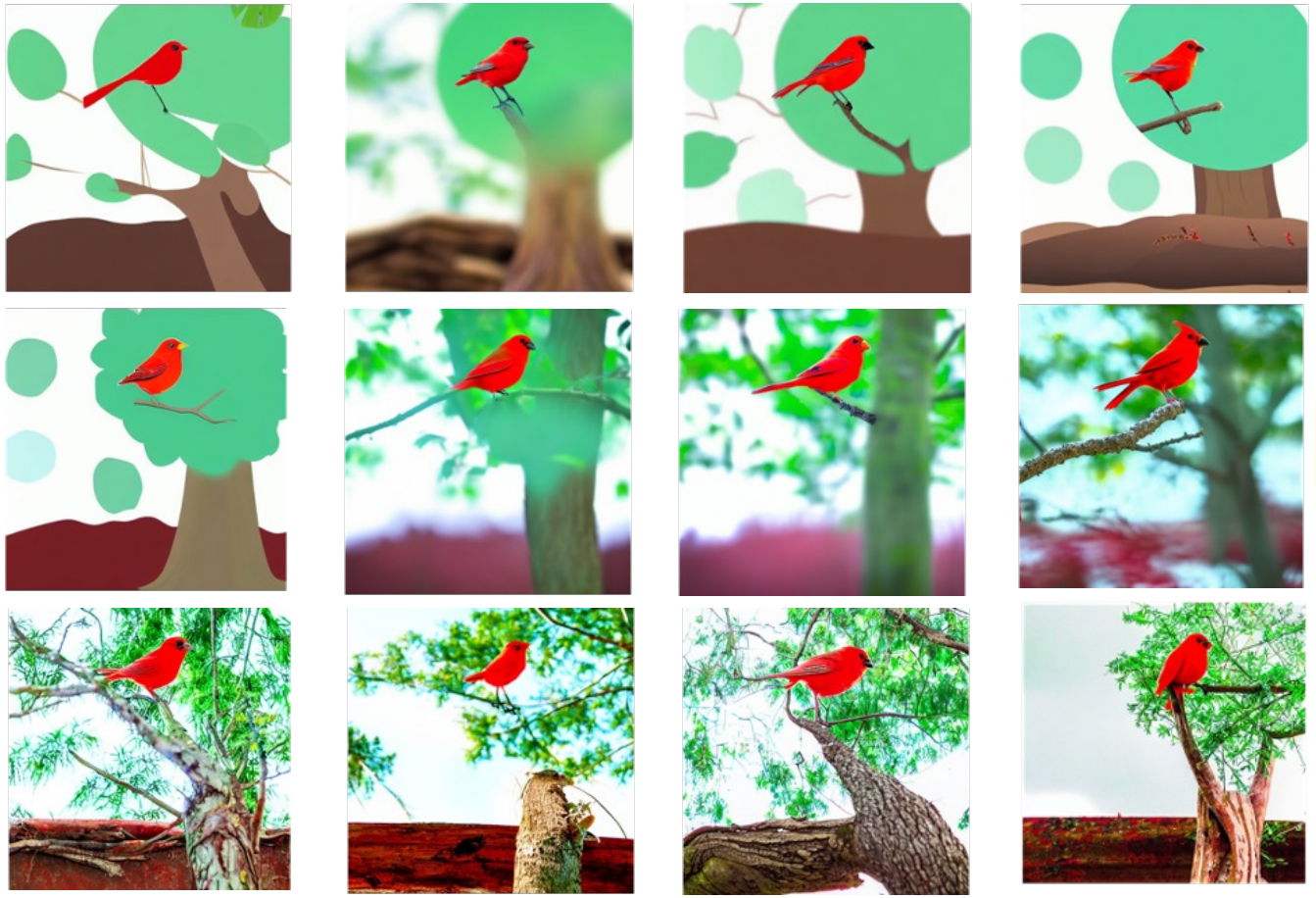
GradOP (Ours)

Text Prompt: "a photo of a red bird in a tree"

Target Subspace

Text Prompt:
"a photo of a red bird in a tree"

Reference Painting



SDEdit

Iterative Loopback

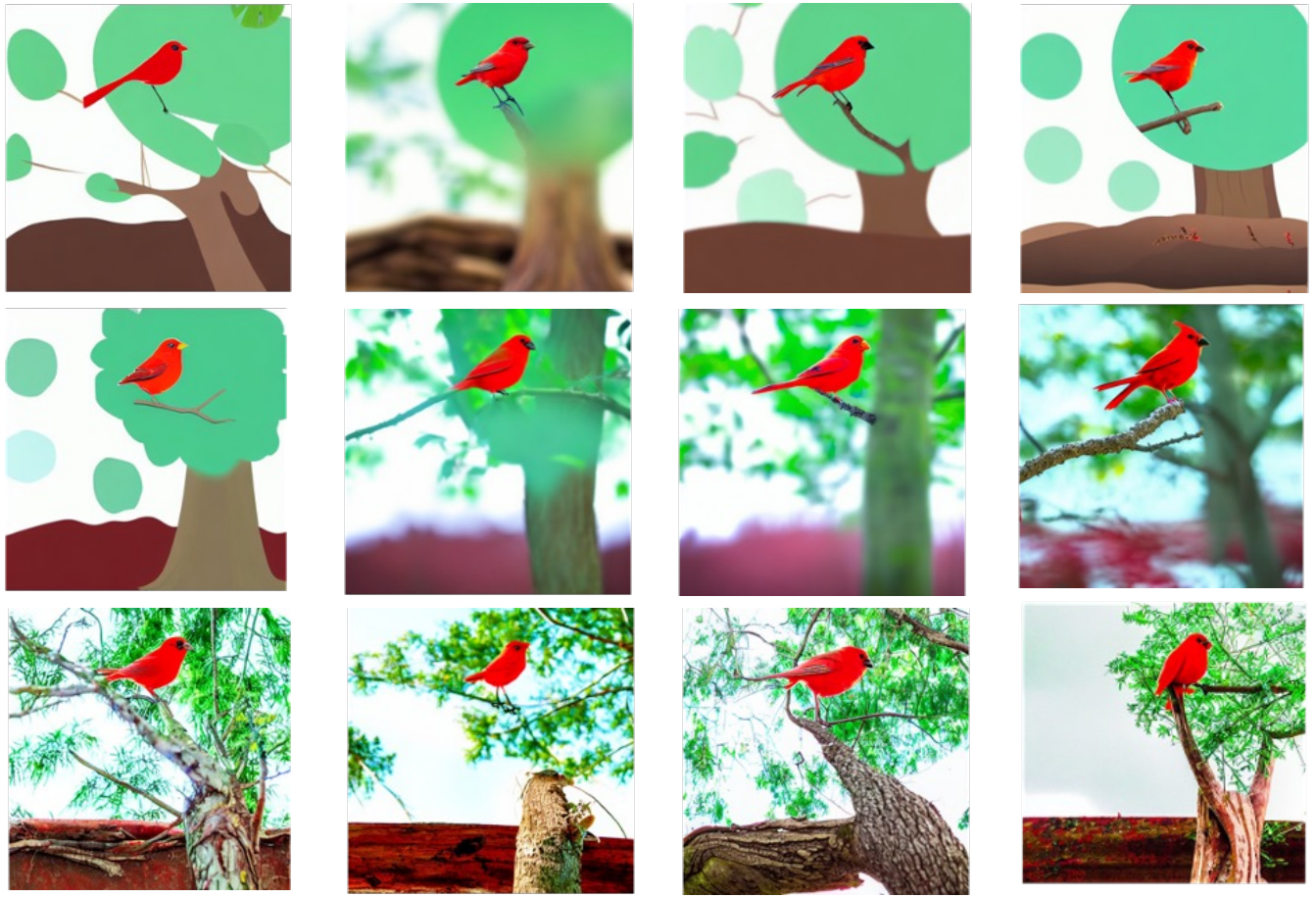
GradOP (Ours)

Text Prompt: "a photo of a red bird in a tree"

Target Subspace

Text Prompt:
"a photo of a red bird in a tree"

Reference Painting



SDEdit

Iterative Loopback

GradOP (Ours)

Text Prompt: "a photo of a blue elephant in a field"

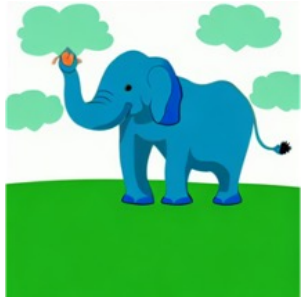



Target Subspace

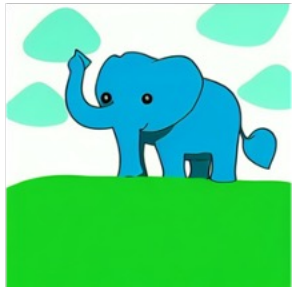
Text Prompt:
"a photo of a blue elephant in a field"

Reference Painting



SDEdit




Iterative Loopback




GradOP (Ours)

Text Prompt: "a photo of a blue elephant in a field"

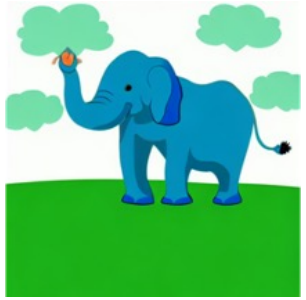



Target Subspace

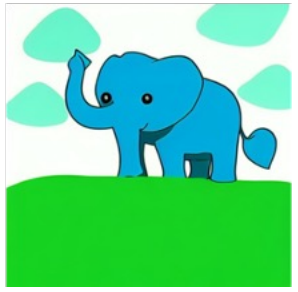
Text Prompt:
"a photo of a blue elephant in a field"

Reference Painting



SDEdit




Iterative Loopback




GradOP (Ours)

Text Prompt: "a photo of a blue elephant in a field"

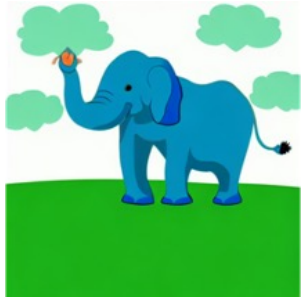



Target Subspace

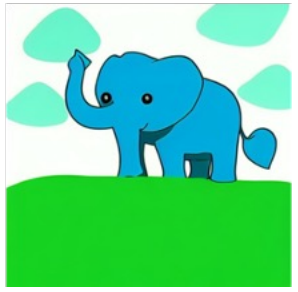
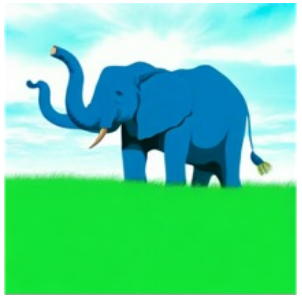
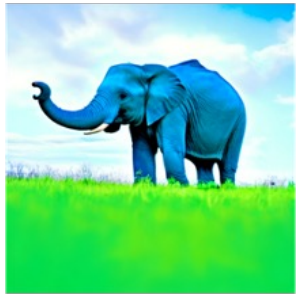
Text Prompt:
"a photo of a blue elephant in a field"

Reference Painting



SDEdit


Iterative Loopback




GradOP (Ours)

Text Prompt: "a photo of castle by the lake"

Target Subspace

$t_0 = 0.7$ $t_0 = 0.75$ $t_0 = 0.8$ $t_0 = 0.85$ $t_0 = 0.9$ $t_0 = 0.95$ $t_0 = 0.99$

Reference Painting

SDEdit

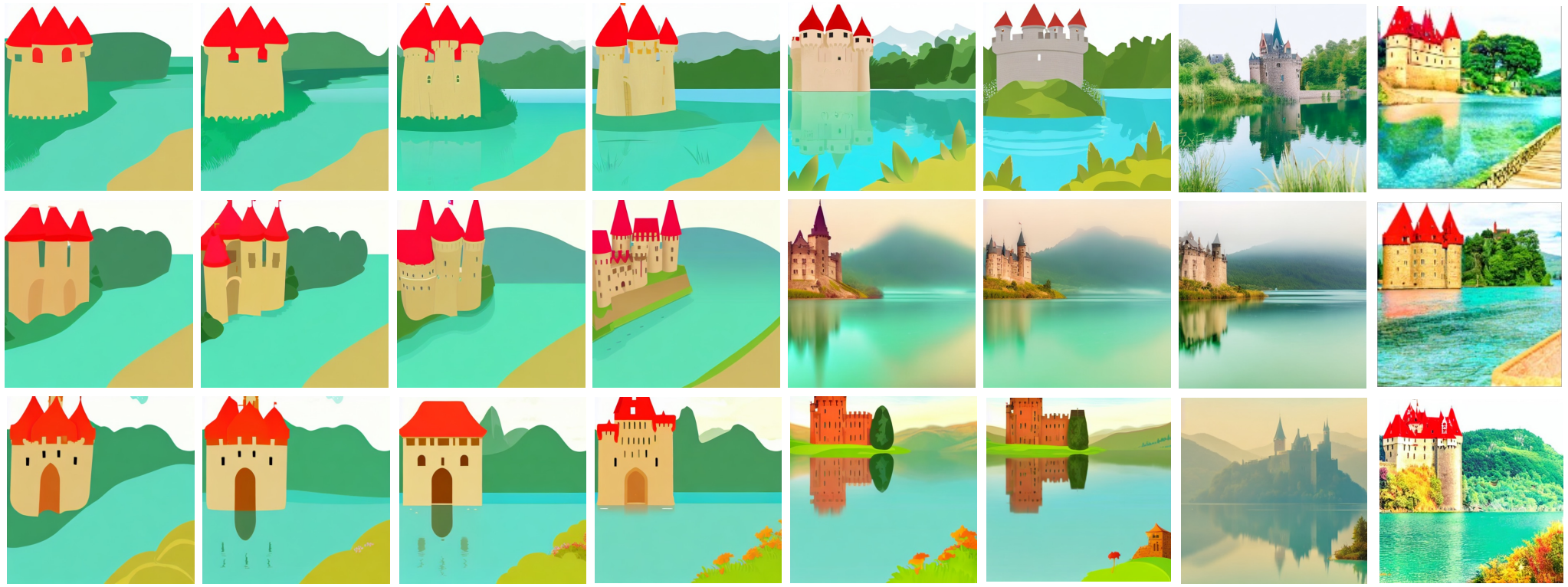
Ours

Text Prompt: "a photo of castle by the lake"

Target Subspace

$t_0 = 0.7$ $t_0 = 0.75$ $t_0 = 0.8$ $t_0 = 0.85$ $t_0 = 0.9$ $t_0 = 0.95$ $t_0 = 0.99$

Reference Painting



SDEdit

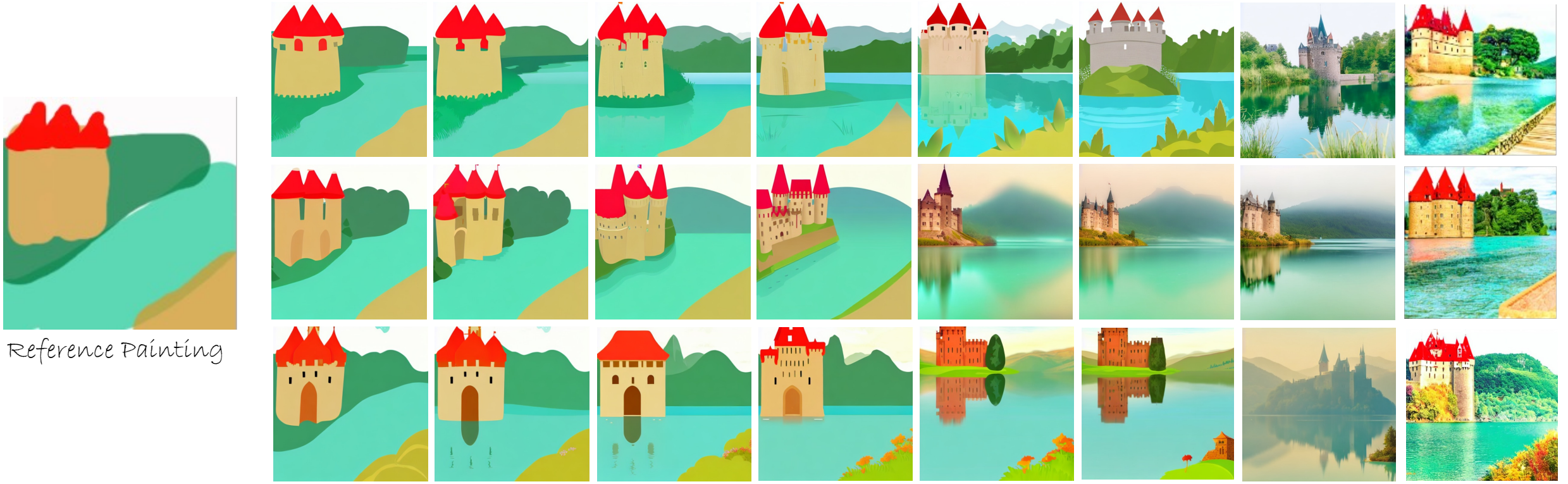
Ours

Faithful but unrealistic

Text Prompt: "a photo of castle by the lake"

Target Subspace

$t_0 = 0.7$ $t_0 = 0.75$ $t_0 = 0.8$ $t_0 = 0.85$ $t_0 = 0.9$ $t_0 = 0.95$ $t_0 = 0.99$



Reference Painting

SDEdit

Ours

Faithful but unrealistic

Realistic but unfaithful

Text Prompt: "a photo of castle by the lake"

Target Subspace

$t_0 = 0.7$ $t_0 = 0.75$ $t_0 = 0.8$ $t_0 = 0.85$ $t_0 = 0.9$ $t_0 = 0.95$ $t_0 = 0.99$

Reference Painting

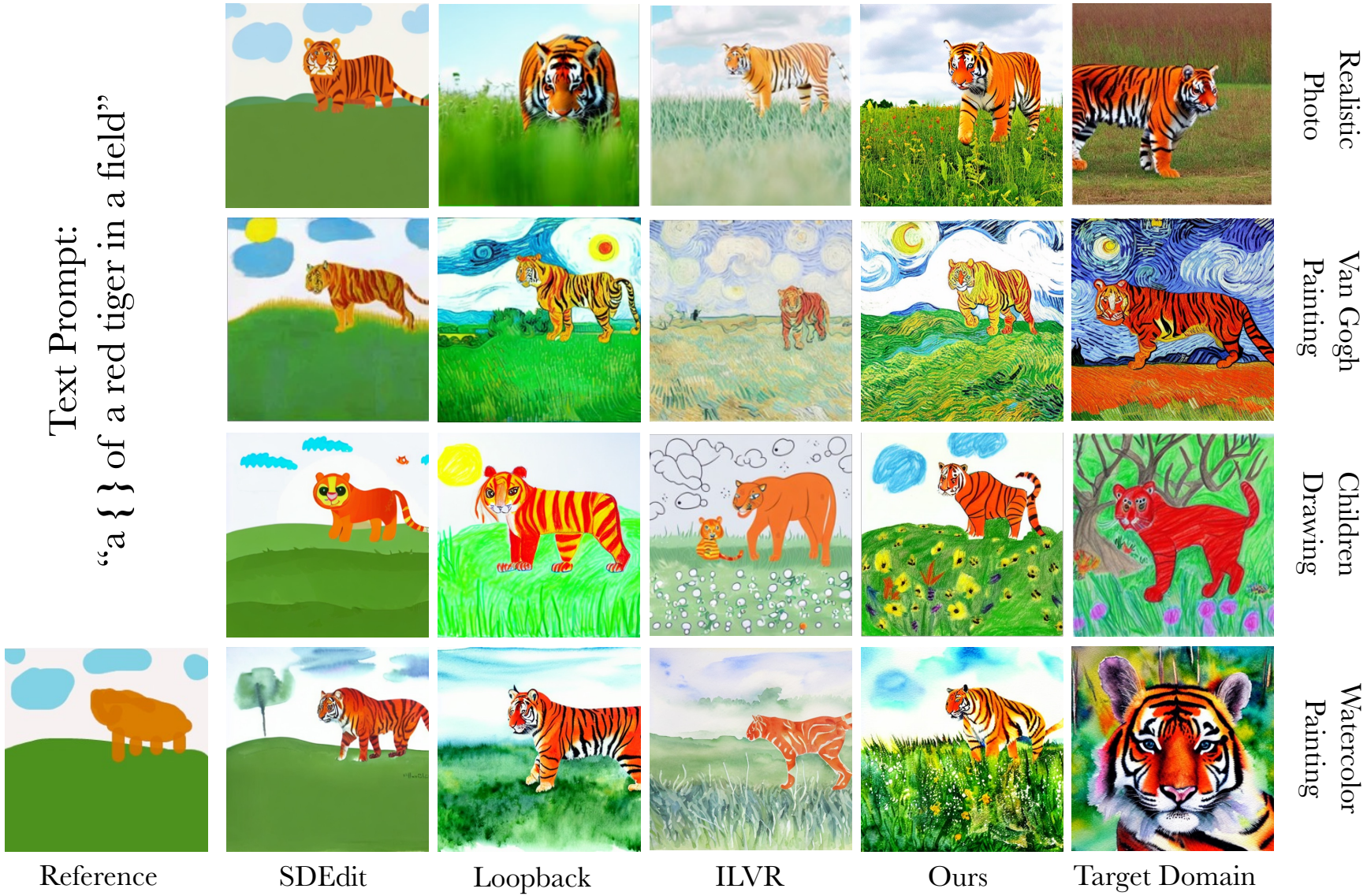
SDEdit

Faithful but unrealistic

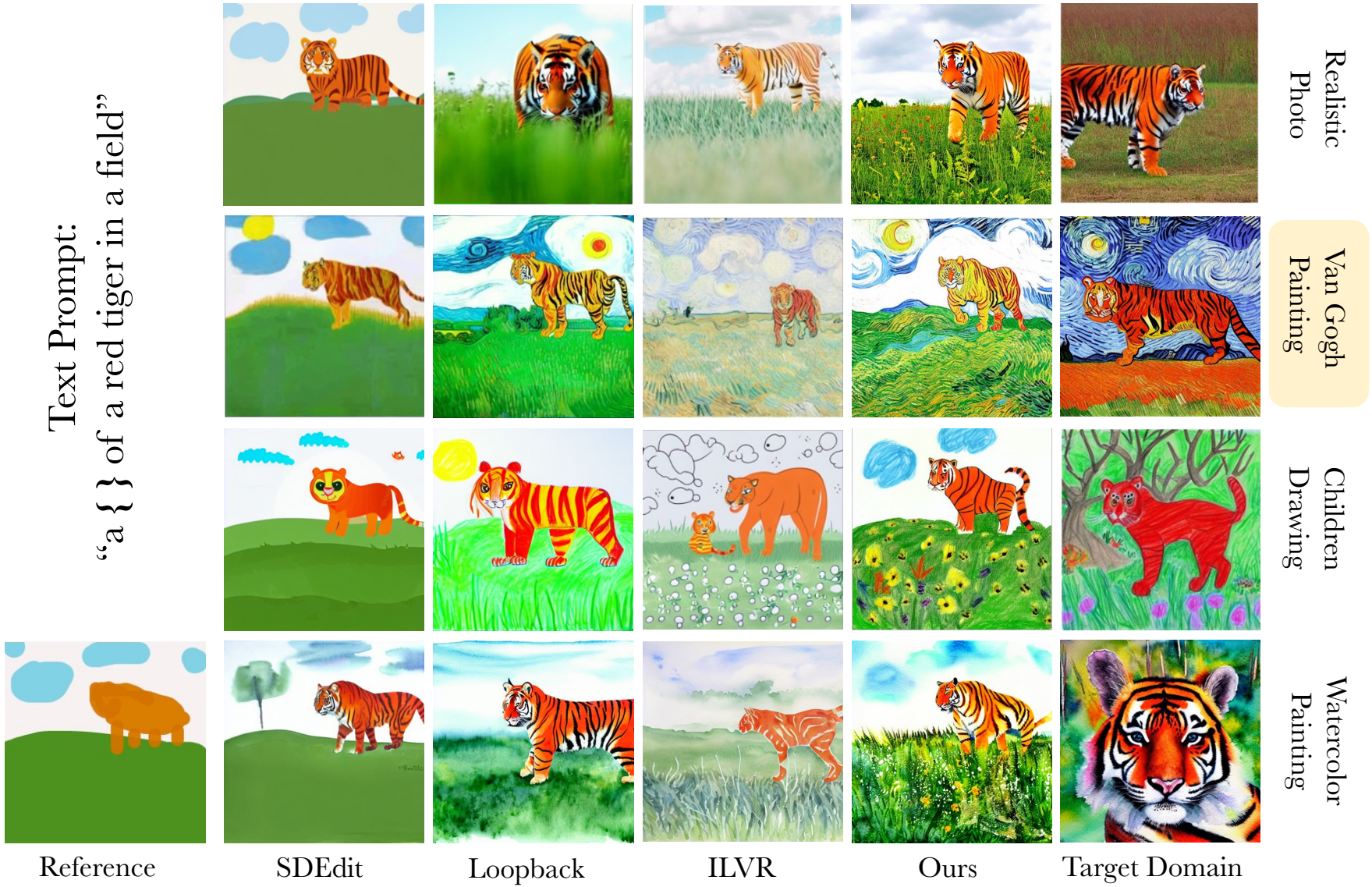
Realistic but unfaithful

Ours

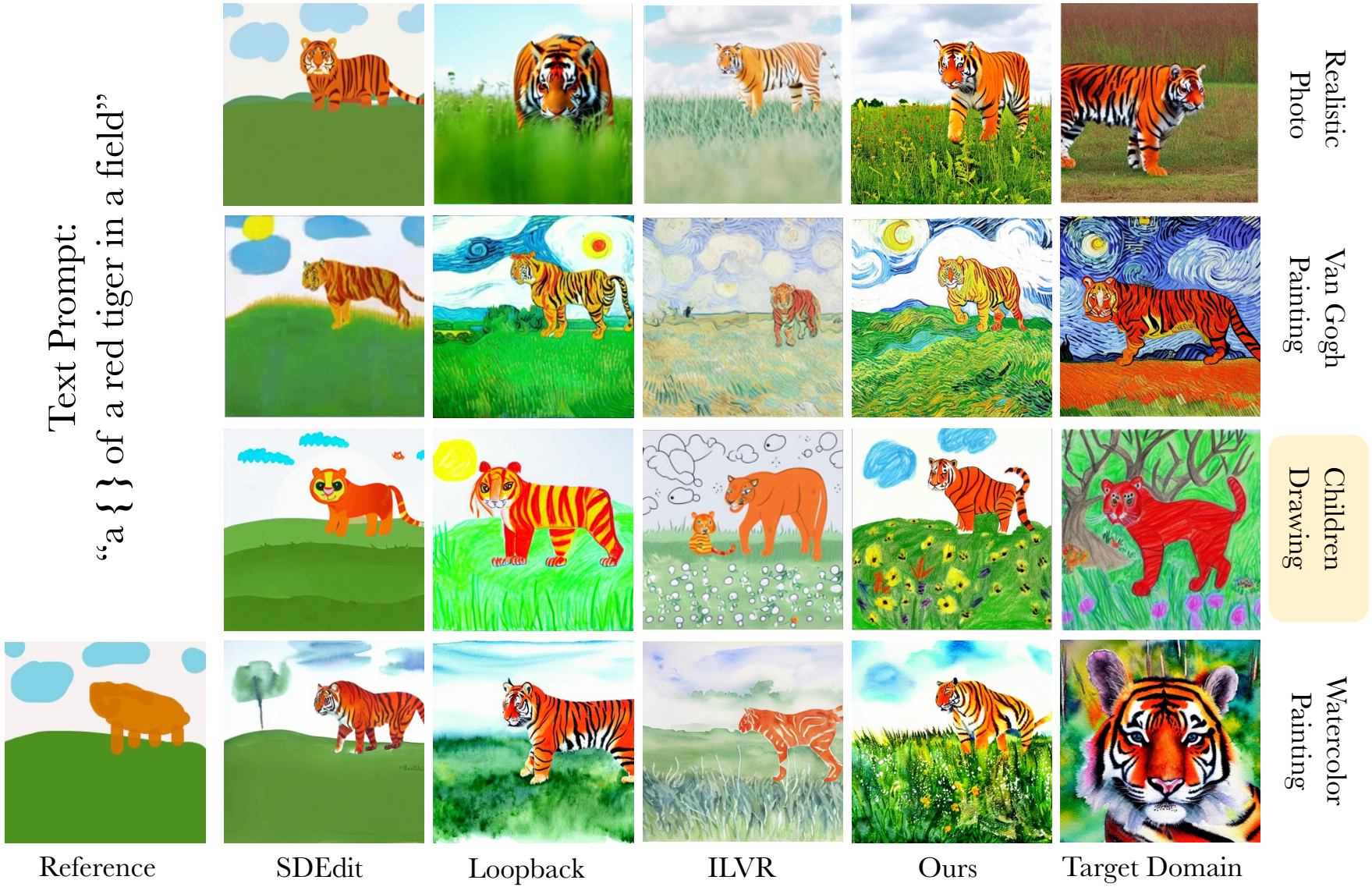
Generalizability across Target Domains



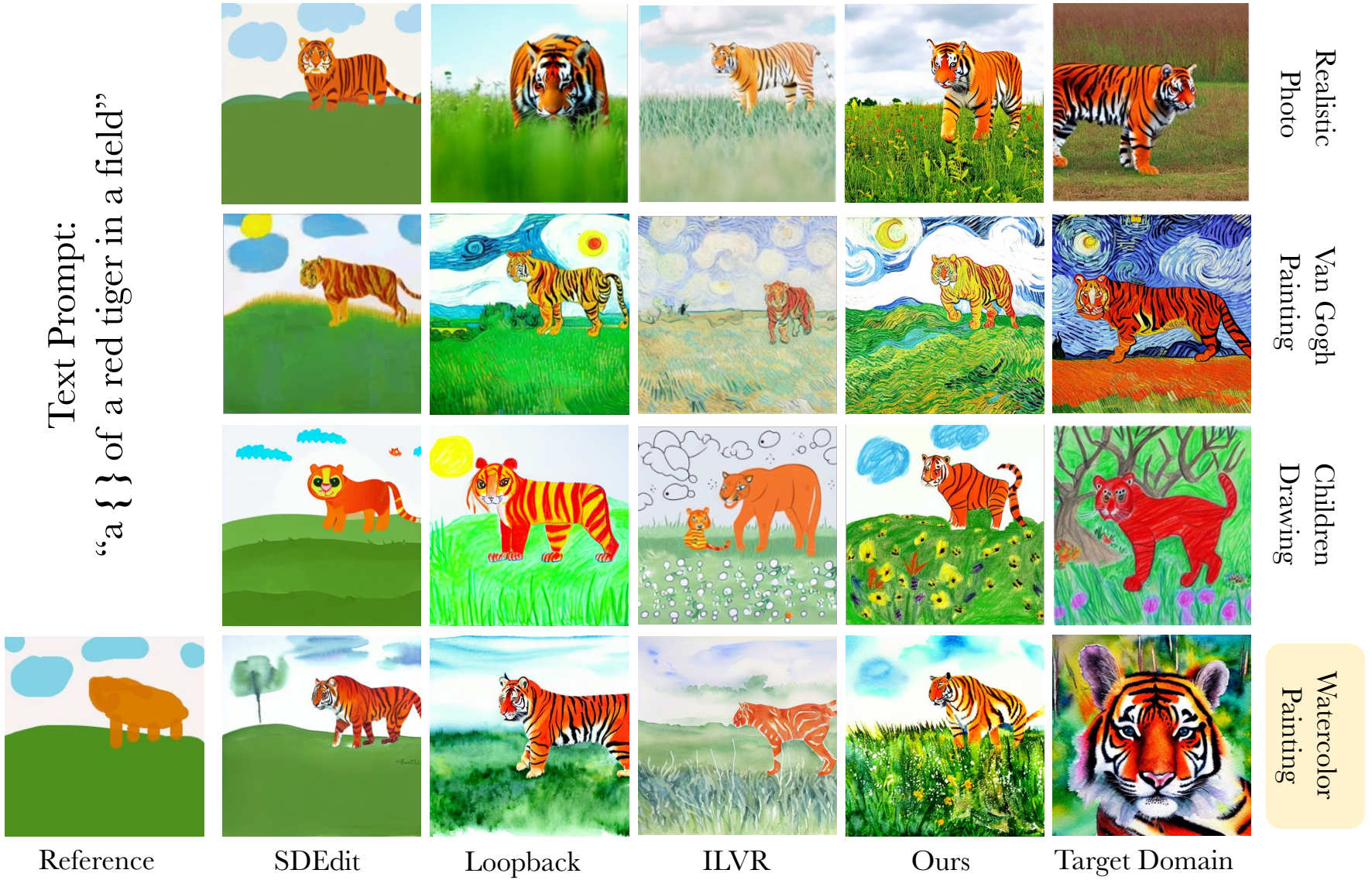
Generalizability across Target Domains



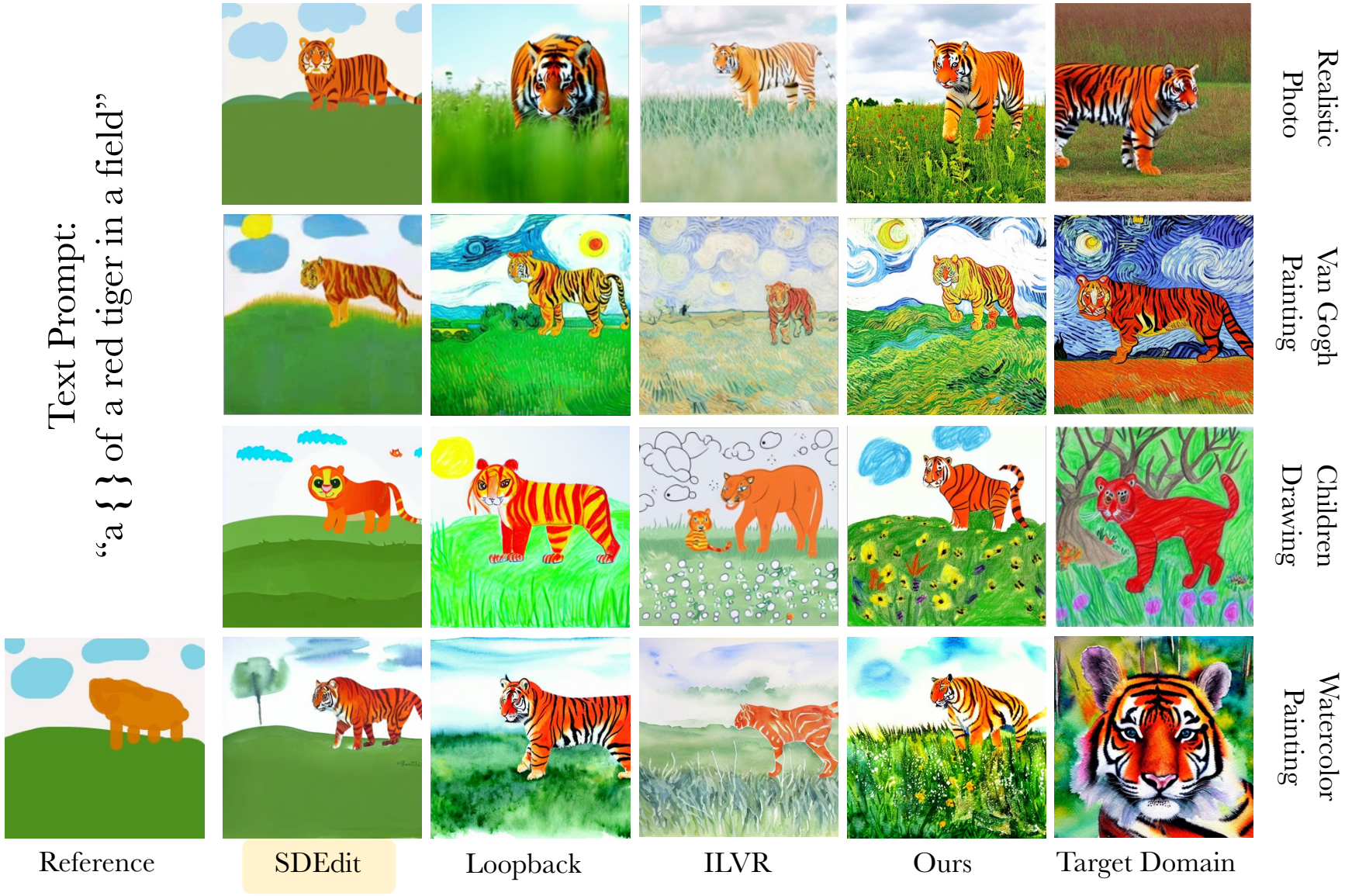
Generalizability across Target Domains



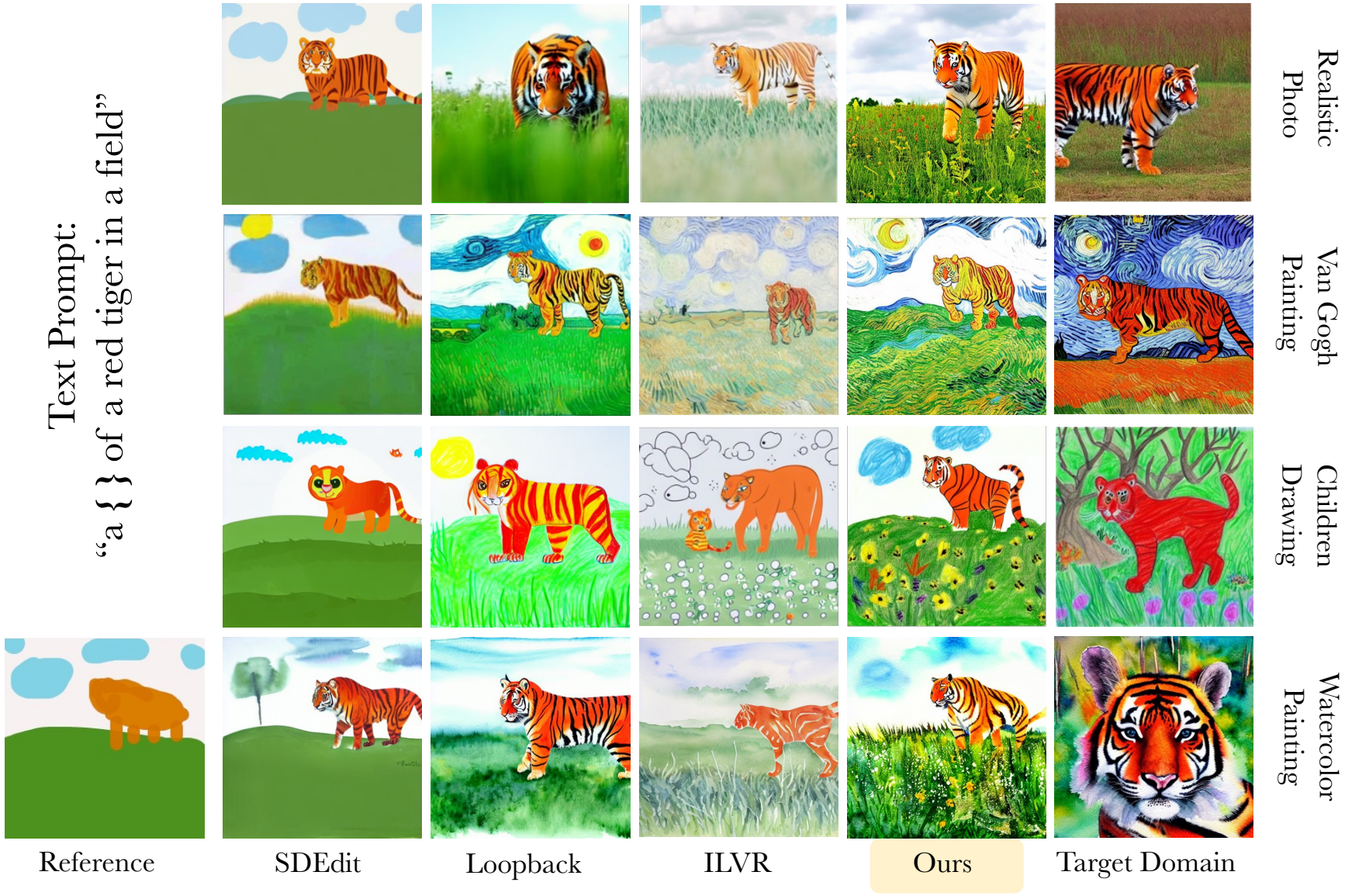
Generalizability across Target Domains



Generalizability across Target Domains



Generalizability across Target Domains



Controlling Semantics of Different Painting Regions

Text Prompt:

"a fantasy landscape, trending on artstation"

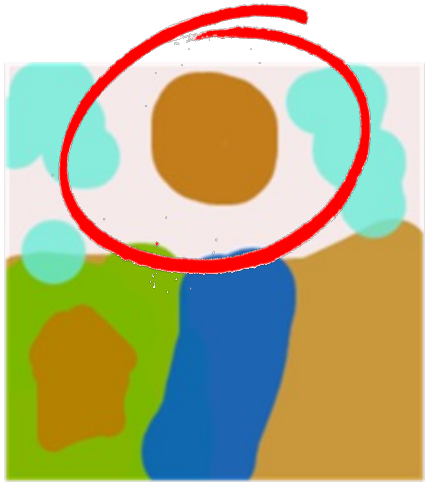
Reference Painting

Semantics of different painting regions might not accurately reflect user's intent.

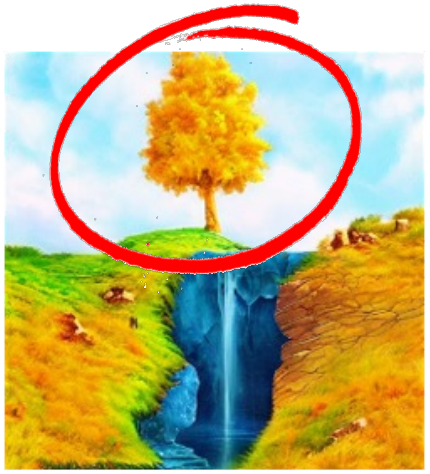
Controlling Semantics of Different Painting Regions

Text Prompt:

"a fantasy landscape, trending on artstation"



Reference Painting



Semantics of different painting regions might not accurately reflect user's intent.

Controlling Semantics of Different Painting Regions

Text Prompt:

"a fantasy landscape, trending on artstation"

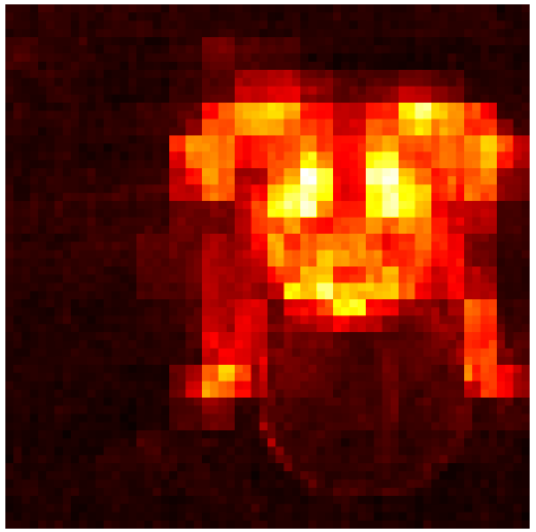
Reference Painting

Some regions might be entirely omitted if the model does not understand that it corresponds to a separate semantic entity

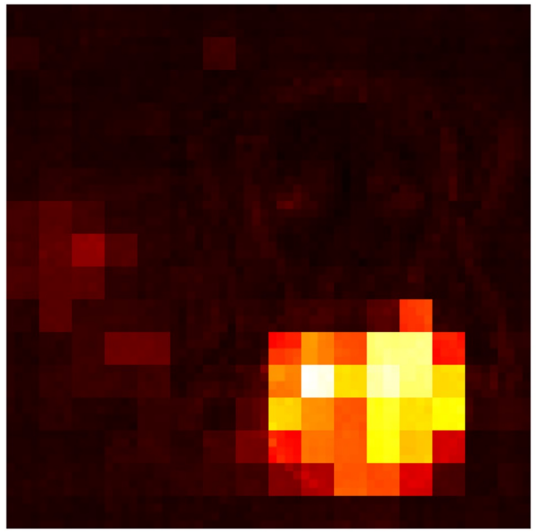
Controlling Semantics of Different Painting Regions

Text Prompt:
"a dog with a ball"

"dog"



"ball"



Average cross-attention maps during reverse diffusion process

Cross-Attention maps show high overlap with semantic segmentation of different entities

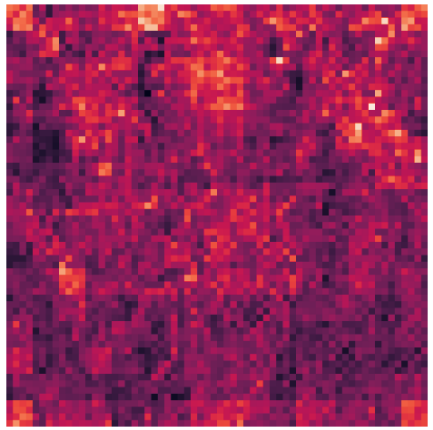
Reference Painting

Binary Mask for
Painting Region

Original
Cross-attention Map

Reference Painting

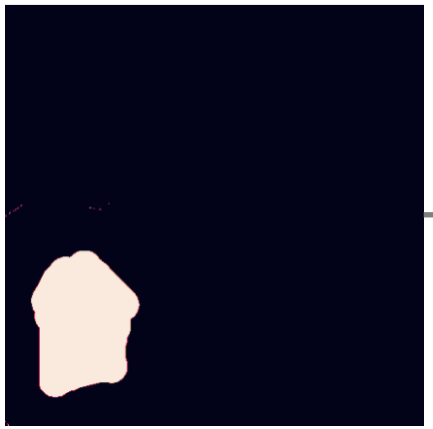
$\mathcal{B}_i$: “hut”



$\mathcal{A}_i^t$

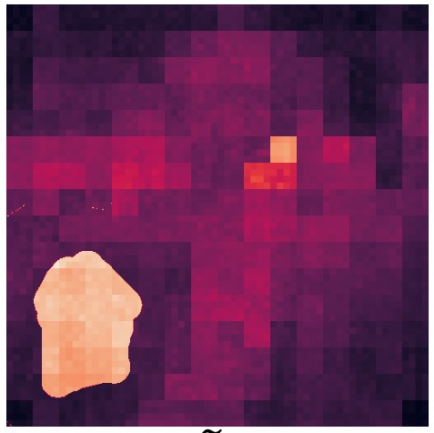
Reference Painting

Binary Mask for Painting Region



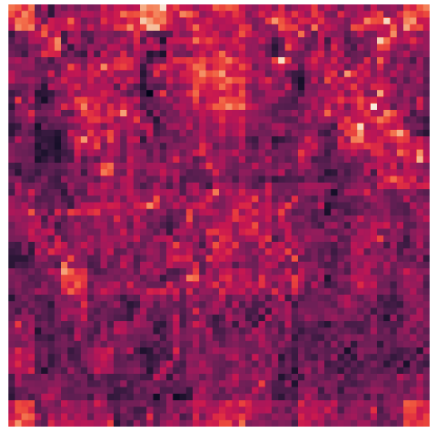
$\mathcal{B}_i$: "hut"

Modified Cross-attention Map



$\tilde{\mathcal{A}}_i^t$

Original Cross-attention Map



$\mathcal{A}_i^t$

Introducing Cross-attention Overlap with Desired Painting Regions

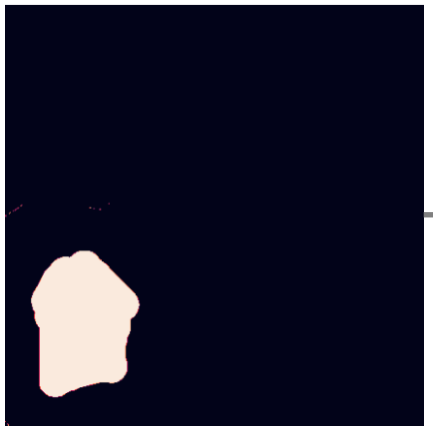
$$\tilde{\mathcal{A}}_t^i = w_i \left[(1 - \kappa_t)\, \mathcal{A}_t^i + \kappa_t \, \frac{\mathcal{B}_i}{\|\mathcal{B}_i\|_F} \, \|\mathcal{A}_t^i\|_F \right]$$

Binary Mask for Painting Region

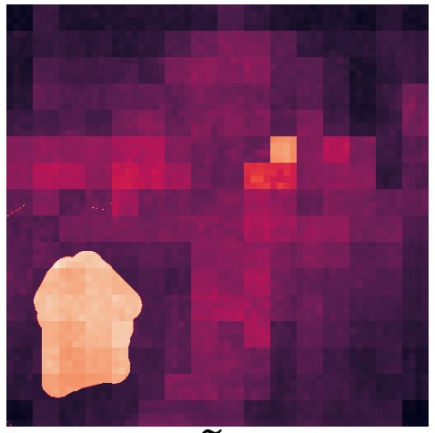
Modified Cross-attention Map

Original Cross-attention Map

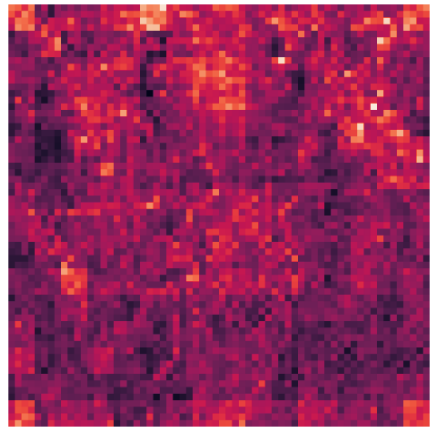
Reference Painting



$\mathcal{B}_i$: "hut"



$\tilde{\mathcal{A}}_i^t$



$\mathcal{A}_i^t$

Generated Image

Introducing Cross-attention Overlap with Desired Painting Regions

$$\tilde{\mathcal{A}}_t^i = w_i \left[(1 - \kappa_t)\, \mathcal{A}_t^i + \kappa_t \frac{\mathcal{B}_i}{\|\mathcal{B}_i\|_F} \|\mathcal{A}_t^i\|_F \right]$$

Reference Painting

Generation Outputs – w/o semantic control

Reference Painting

Generation Outputs – w/o semantic control

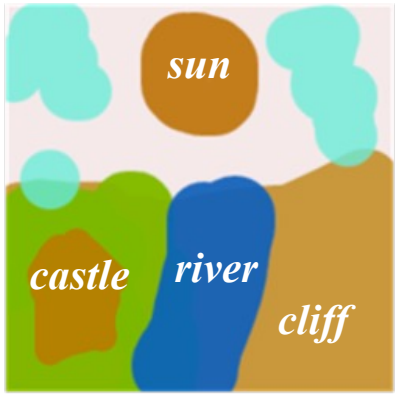
Semantic Guide

Generation Outputs – with semantic control

Reference Painting

Generation Outputs – w/o semantic control

Semantic Guide



Generation Outputs – with semantic control

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thanks for listening!

Project Page and Online Demo
https://1jsingh.github.io/gradop

† Australian National University

★ AUSTRALIAN CENTRE FOR ROBOTIC VISION

Jaskirat Singh†

Stephen Gould†*

Liang Zheng†*

Poster Tag: TUE-PM-179