RUTGERS THE STATE UNIVERSITY OF NEW JERSEY Snap Inc.

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

SINE: SINgle Image Editing with Text-to-Image Diffusion Models

Zhixing Zhang[1] Ligong Han[1] Arnab Ghosh[2] Dimitris Metaxas[1] Jian Ren[2]

[1] Rutgers University [2] Snap Inc.

RUTGERS THE STATE UNIVERSITY OF NEW JERSEY Snap Inc.

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

SINE: SINgle Image Editing with Text-to-Image Diffusion Models

Zhixing Zhang[1] Ligong Han[1] Arnab Ghosh[2] Dimitris Metaxas[1] Jian Ren[2]

[1] Rutgers University [2] Snap Inc.

Motivation

- Pre-trained diffusion model for real image editing



"Oil painting of a lady in the style of Pierre-Auguste Renoir" ($H = 512,\ W = 512$)

Source IMG DreamBooth Textual-Inversion



"A sculpture of a lady in the style of Michelangelo" ($H = 768,\ W = 768$)

Source IMG DreamBooth Textual-Inversion

Overfitting

$\Longrightarrow$ *Model-based guidance*

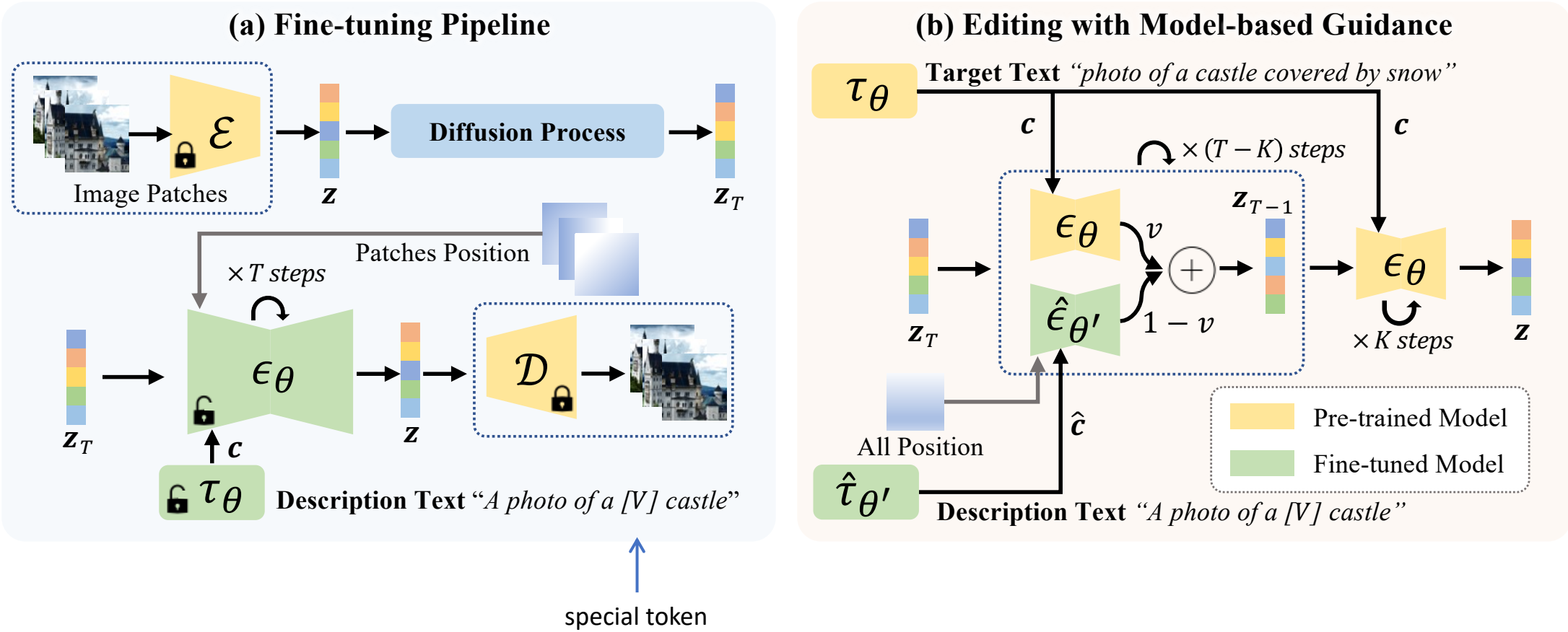
Arbitrary resolution Generation

$\Longrightarrow$ Patch-based fine-tuning

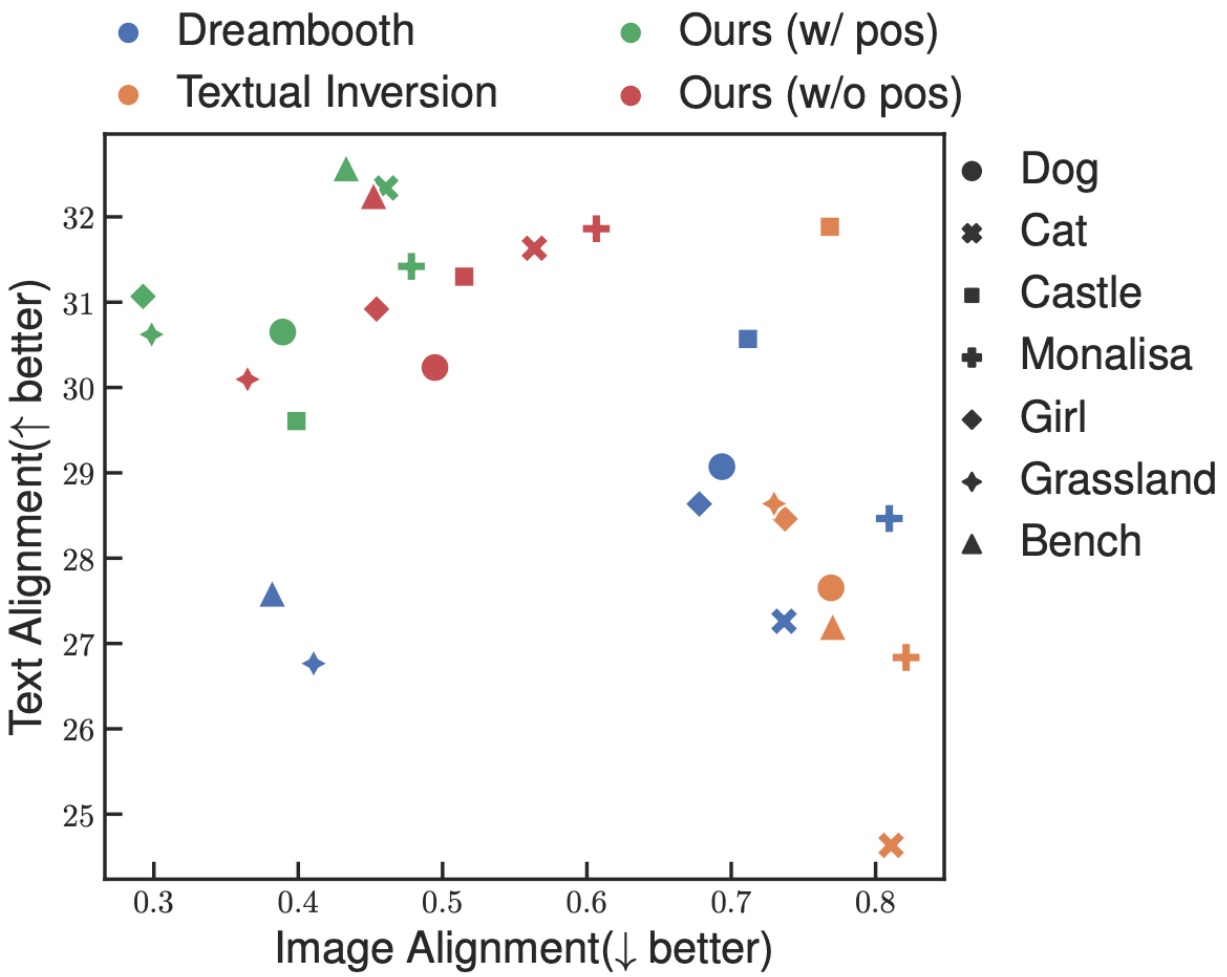
Motivation

- Pre-trained diffusion model for real image editing:
  e.g. Textual-Inversion, DreamBooth
- Overfitting ⟹ *Model-based guidance*
- Arbitrary resolution ⟹ Patch-based fine-tuning

Methods


(a) Fine-tuning Pipeline
Image Patches
Diffusion Process
Patches Position
× T steps
Description Text "A photo of a [V] castle"
special token
(b) Editing with Model-based Guidance
Target Text "photo of a castle covered by snow"
× (T − K) steps
× K steps
All Position
Description Text "A photo of a [V] castle"
Pre-trained Model
Fine-tuned Model

Results


Dreambooth
Textual Inversion
Ours (w/ pos)
Ours (w/o pos)
Dog
Cat
Castle
Monalisa
Girl
Grassland
Bench
Text Alignment(↑ better)
Image Alignment(↓ better)
32
31
30
29
28
27
26
25
0.3
0.4
0.5
0.6
0.7
0.8

Results

- $K = 400, v = 0.7$

Source Image

"A painting of a grassland in the style of Vincent Van Gogh"

"A photo of a grassland in thunderstorm"

Source Image

"A photo of a bench on sand"

"A painting of a bench in the style of Vincent Van Gogh"

Results

- $K = 400, v = 0.7$

Source Image

"A painting of a teddy bear in the style of street painting"

"A cartoon of a teddy bear"

Source Image

"A photo of a tree covered by snow"

"A photo of a tree on fire"

Results

- $K = 400, v = 0.7$

Source Image

"A sculpture of a horse"

"A cartoon of a horse"

Source Image

"modern art painting of a vase"

"painting of a vase in the style of Vincent Van Gogh"

Results

- $K = 400, v = 0.7$

Source Image

"A painting of a girl in the style of Vincent Van Gogh"

"A painting of a girl in the style of Leonardo da Vinci"

Source Image

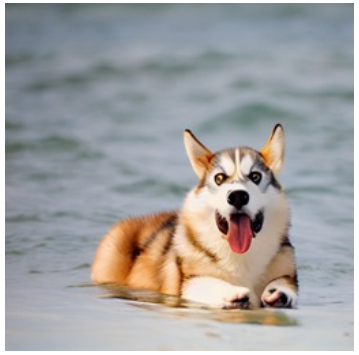
"A photo of a building covered by snow"

"A photo of a cyberpunk building"

Results

- $K = 400, v = 0.7$

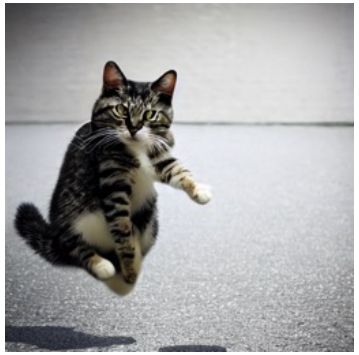
Source Image



"A photo of a Husky in ocean"

"A cartoon of a dog on grass"

Source Image



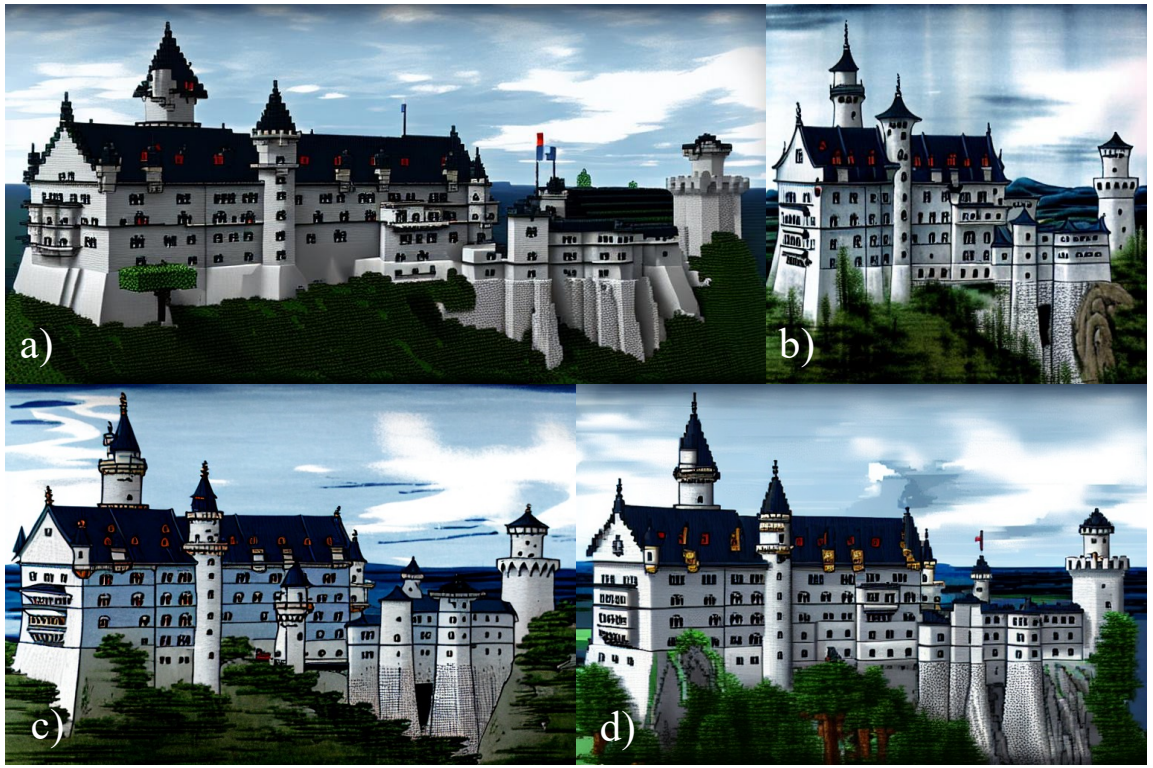
"A photo of a jumping cat"

"A photo of a cat wearing a red hat"

More editing tasks: Different resolutions

- Synthesize images w/ various resolutions

Source Image

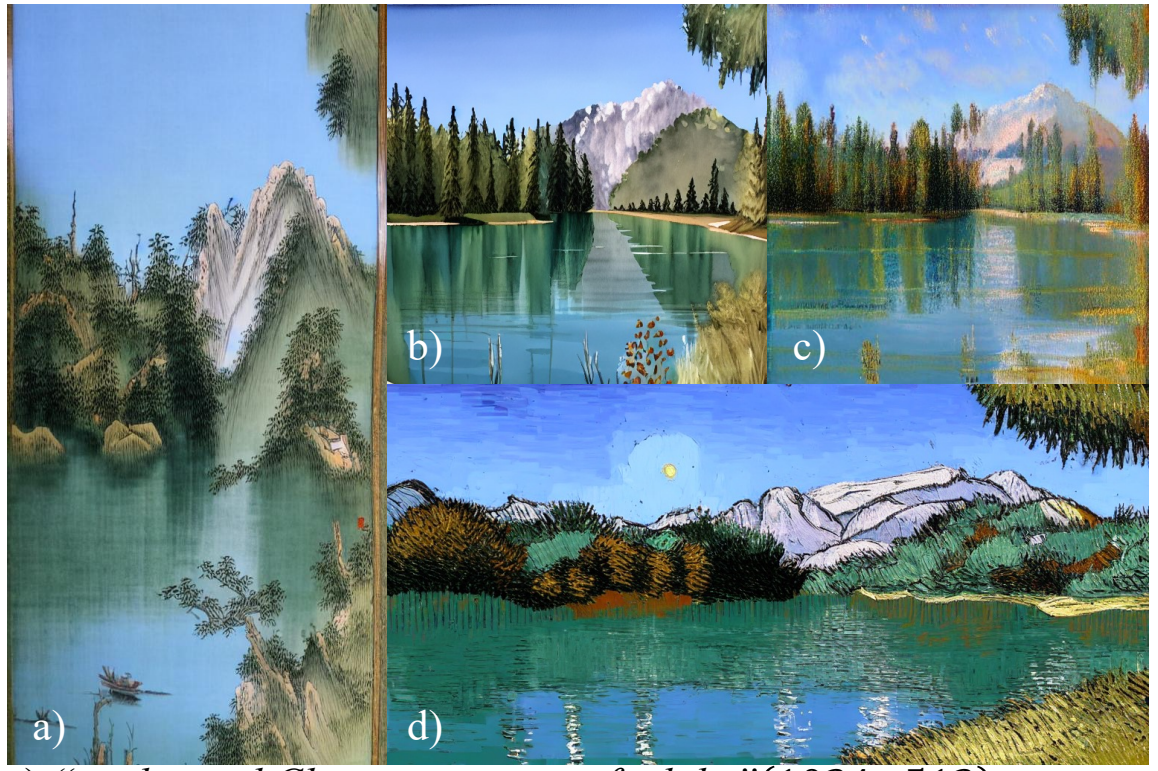


a) *"a castle in the style of Minecraft"* (512×1024)
b) *"a Chinese painting of a castle"*(512×512)
c) *"painting of a castle in the style of Hokusai"*(512×768)
d) *"pixel art of a castle"*(512×768)

More editing tasks: Different resolutions

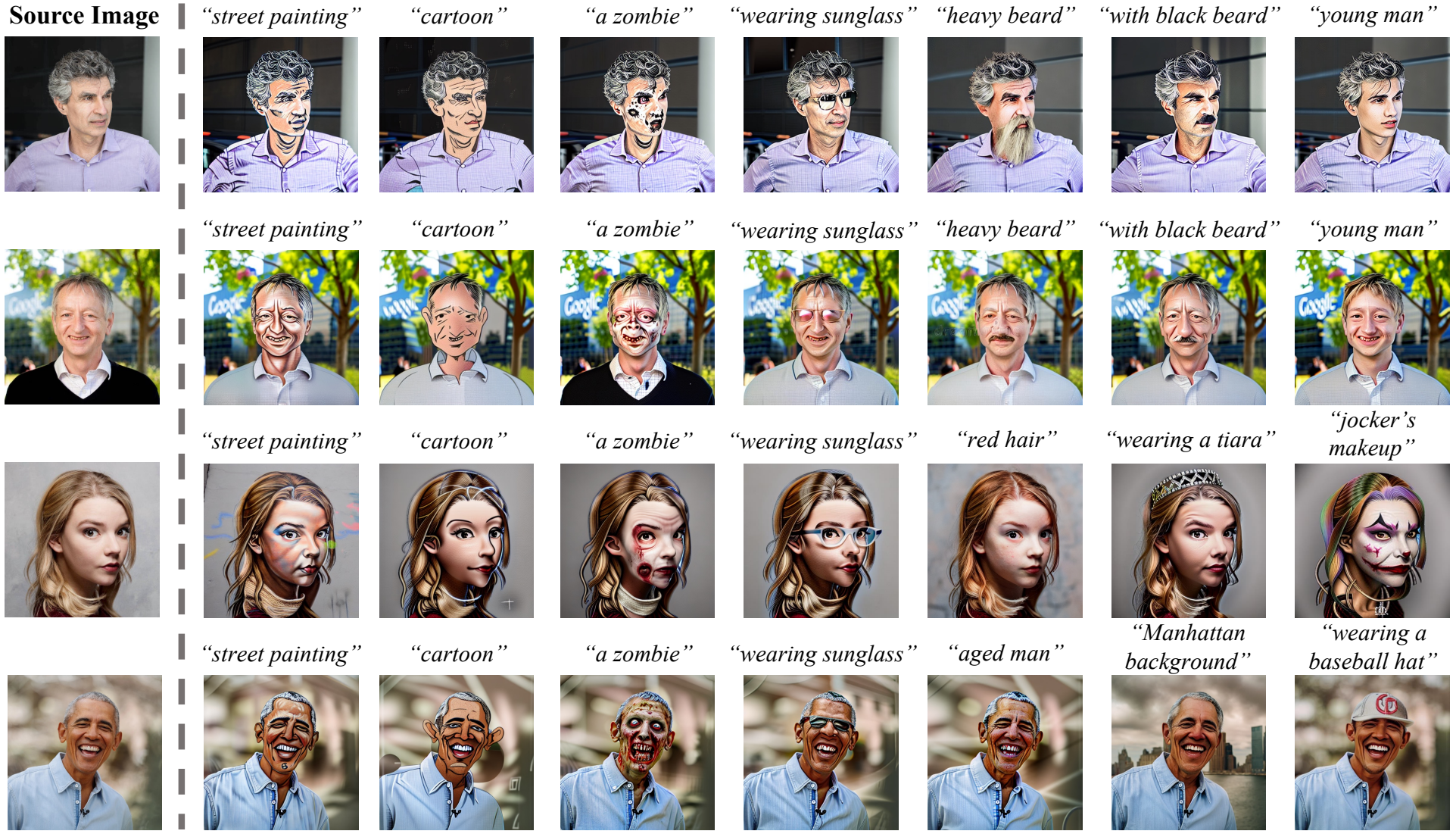
- Synthesize images w/ various resolutions

Source Image



a) *"traditional Chinese painting of a lake"*(1024×512)
b) *"a watercolor painting of a lake of a lake"*(512×512)
c) *"an impressionism painting of a lake"*(512×512)
d) *"painting of a lake ... Vincent Van Gogh"*(512×1024)

More editing tasks: Face editing



More editing tasks: High-resolution

- “A children’s painting of a castle.”
- H=768, W=1024
- $K = 400, v = 0.7$

More editing tasks: High-resolution

- “A painting of a castle in the style of Claude Monet.”
- H=768, W=1024
- $K = 400, v = 0.65$

More editing tasks: High-resolution

- “A desert.”
- H=768, W=1024
- $K = 500, v = 0.8$

More editing tasks: High-resolution

- “A desert.”
- H=768, W=1024
- $K = 500, v = 0.8$

More editing tasks: High-resolution

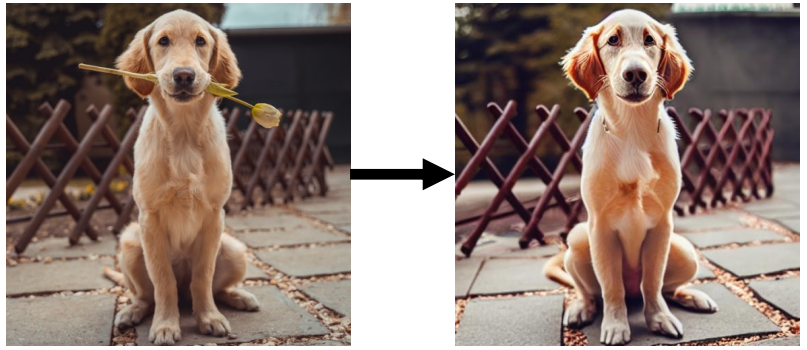
- “A photo of a lake with many sailboats.”
- H=768, W=1024
- $K = 400, v = 0.7$

More editing tasks: High-resolution

- “A watercolor painting of a girl.”
- H=1024, W=768
- $K = 400, v = 0.6$

More editing tasks

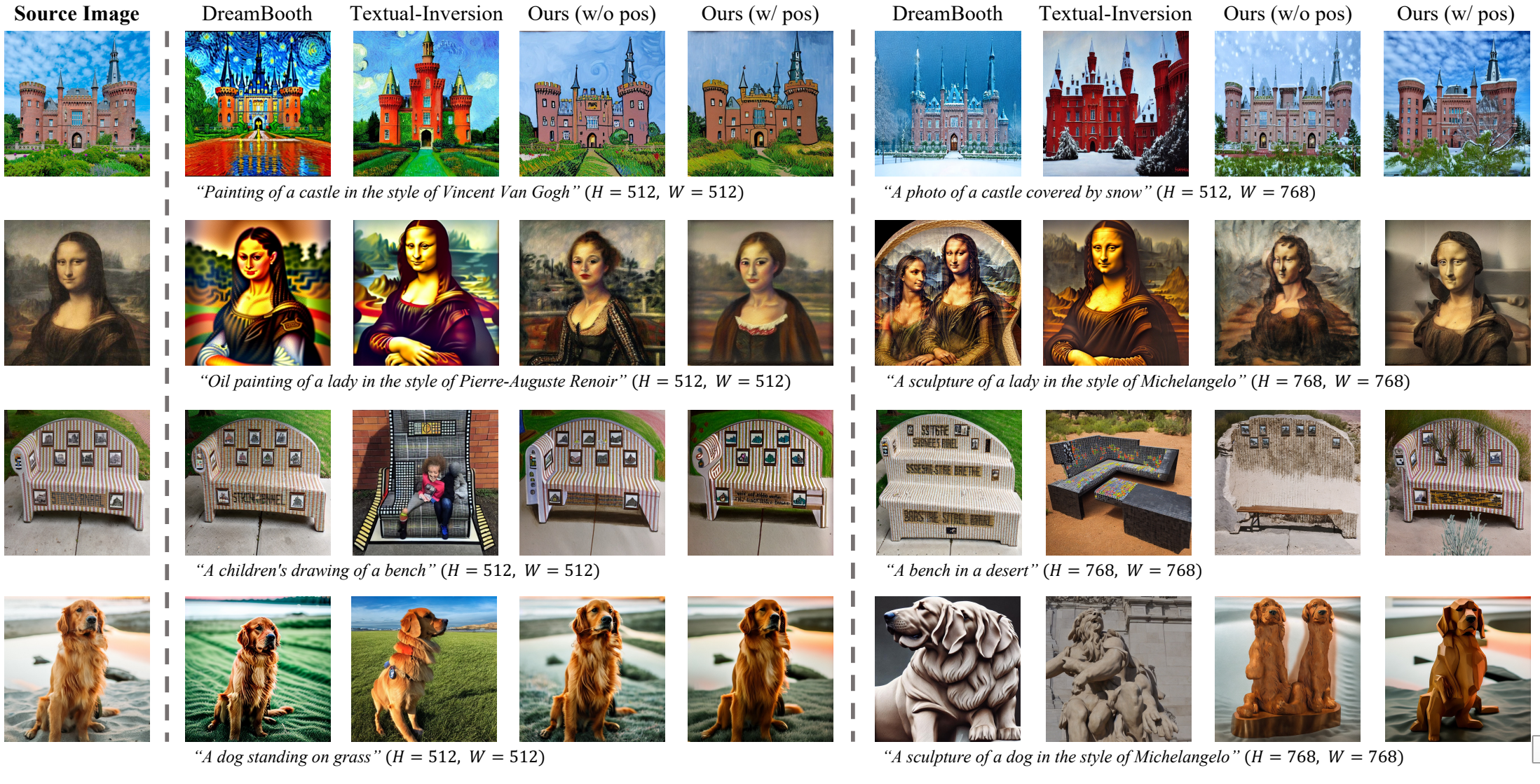
Style Transfer



Content Removal

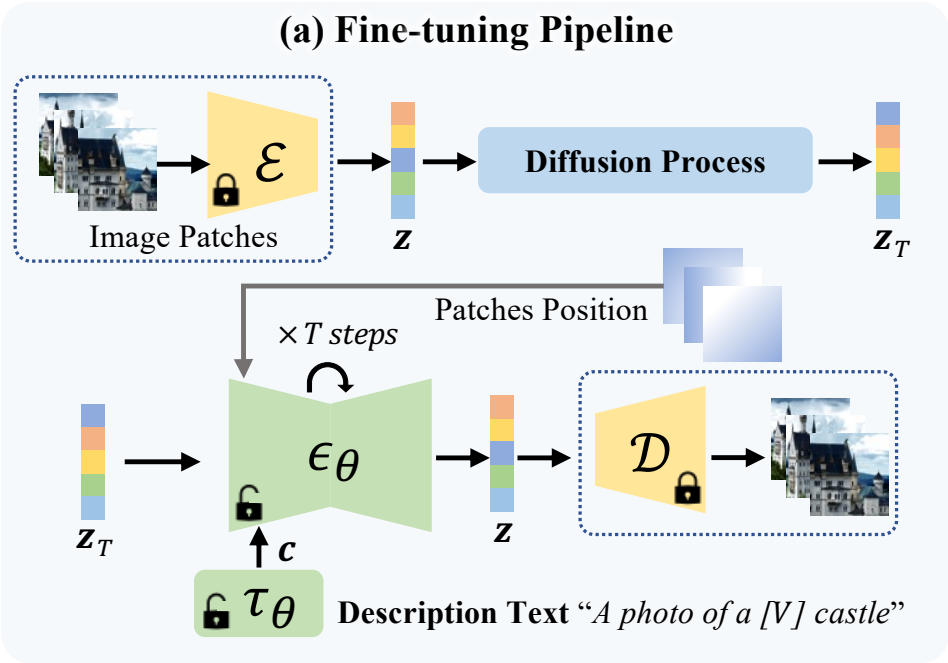
Style Generation

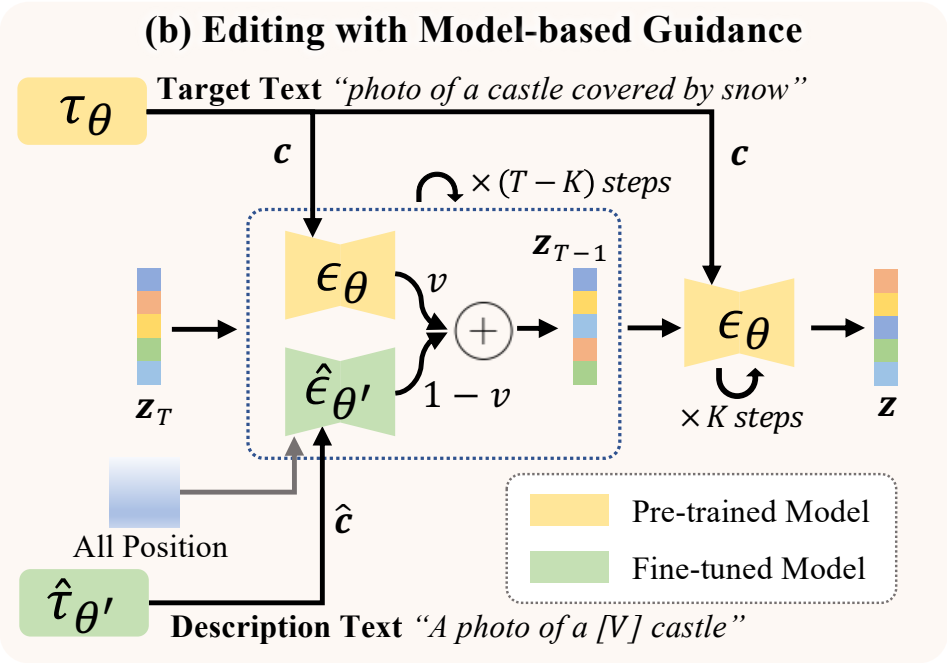
Results: Comparison



Model-based Guidance

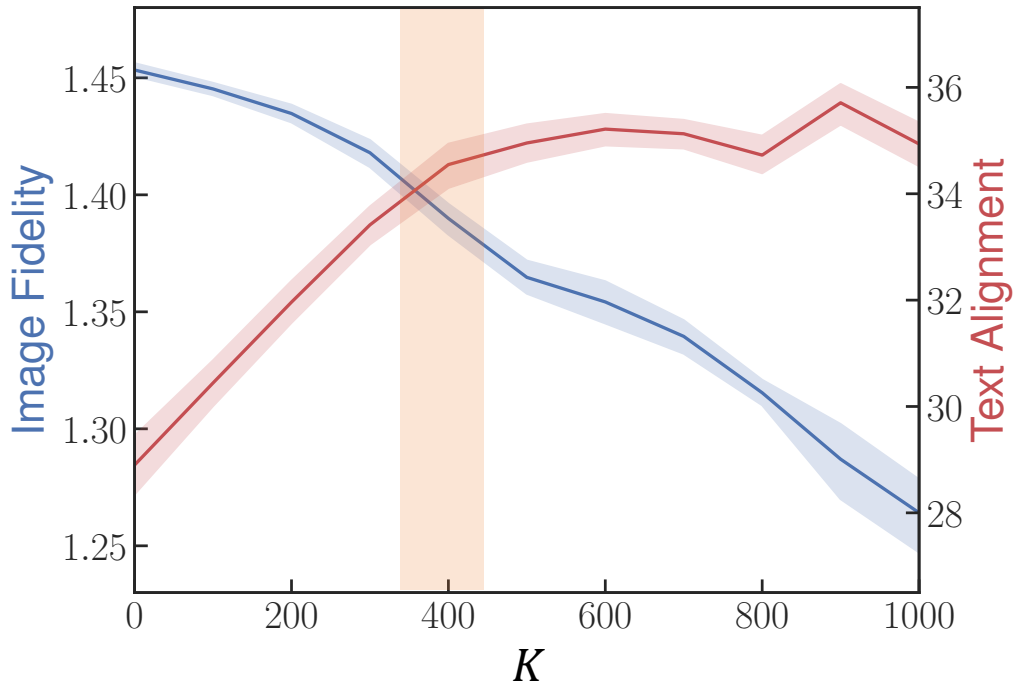
- $\tilde{\epsilon}_\theta(z_t, c) = w\left(v\epsilon_\theta(z_t, c) + (1-v)\hat{\epsilon}_{\theta'}(z_t, \hat{c})\right) + (1-w)\epsilon_\theta(z_t, \emptyset)$ for $t > K$
  - v: strength of editing
  - K: perform guidance from $t = 1000$ to t $= K$
- *decrease* $\downarrow v$ or $\downarrow K \Longrightarrow$ *Increase* $\uparrow$ similarity between *output* sample & *input* image

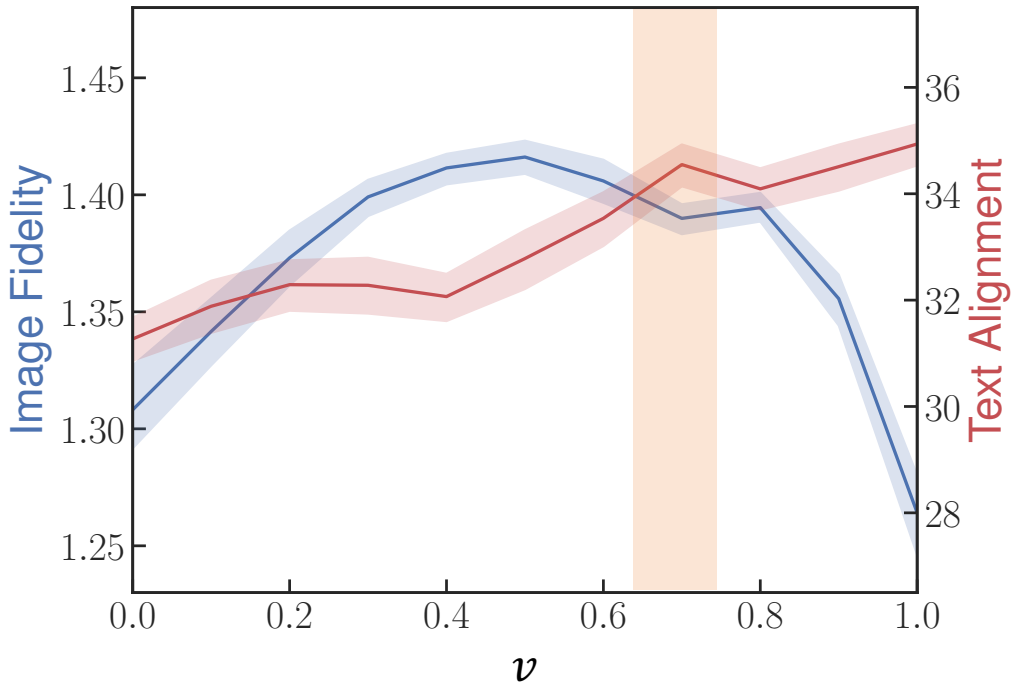




Analysis: Model-based guidance

- Image fidelity: LPIPS score, $1 - \log(\mathcal{L}_{LPIPS})$
- Text alignment: CLIP score ↑, $100 \times \text{cosine}(text, image)$

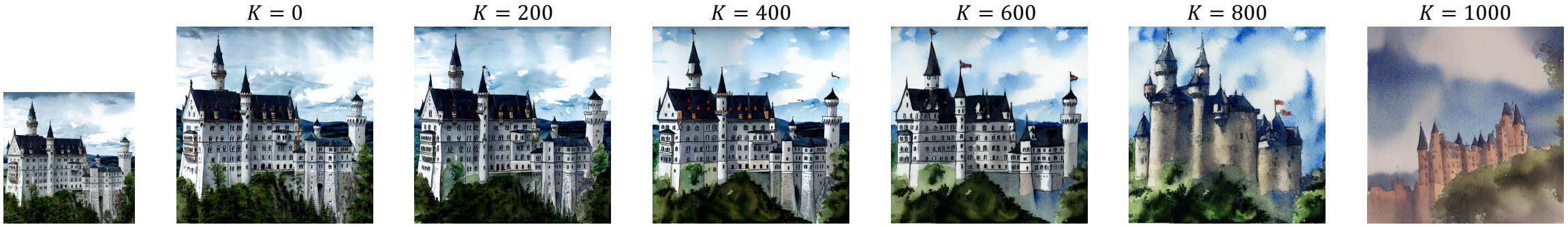




Analysis: Guidance step K

- $v = 0.7$

Source Image "a Chinese painting painting of a castle" ($H = 768,\ W = 768$)



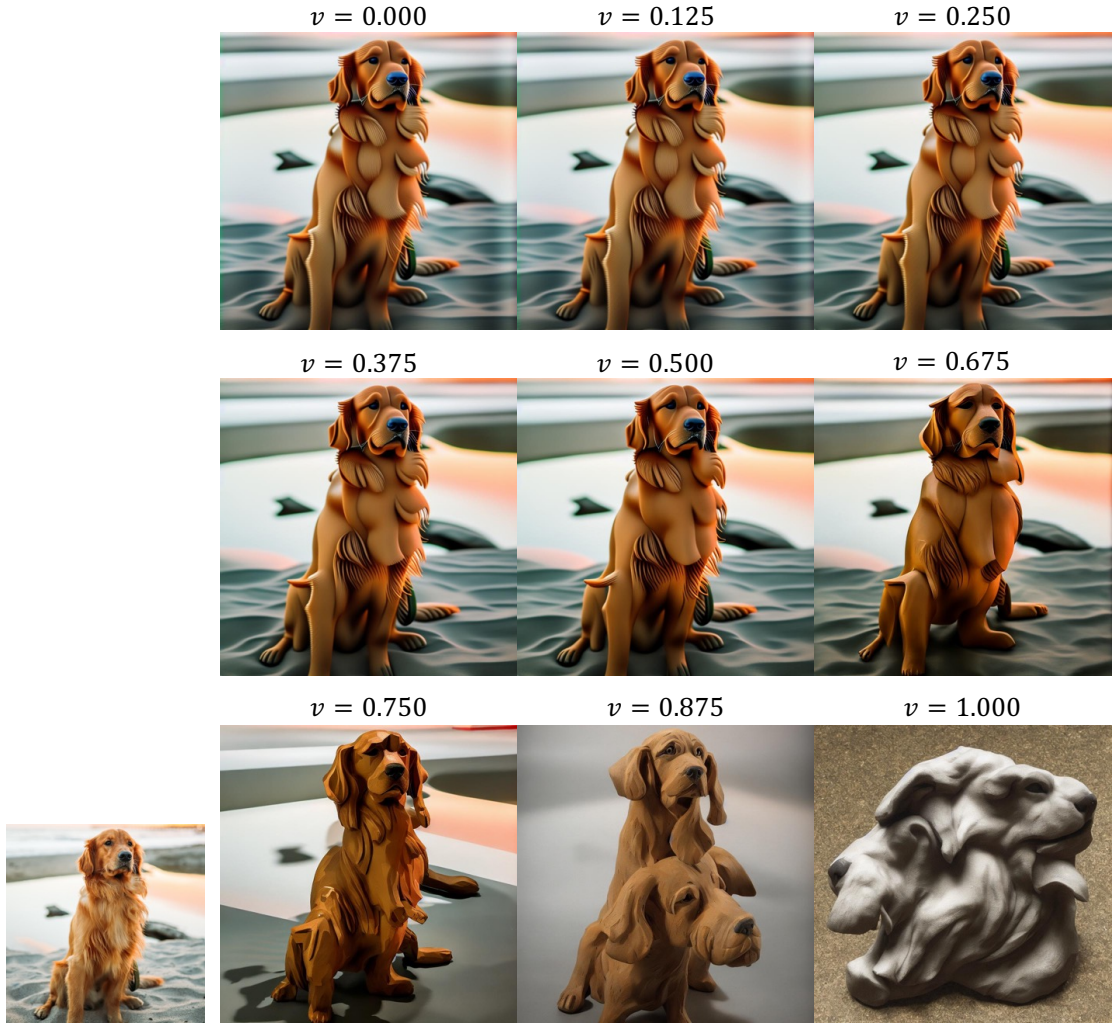
Source Image "a watercolor painting of a castle" ($H = 768,\ W = 768$)

Analysis: Guidance step K

- $v = 0.7$

Analysis: Guidance weight v

- $K = 400$

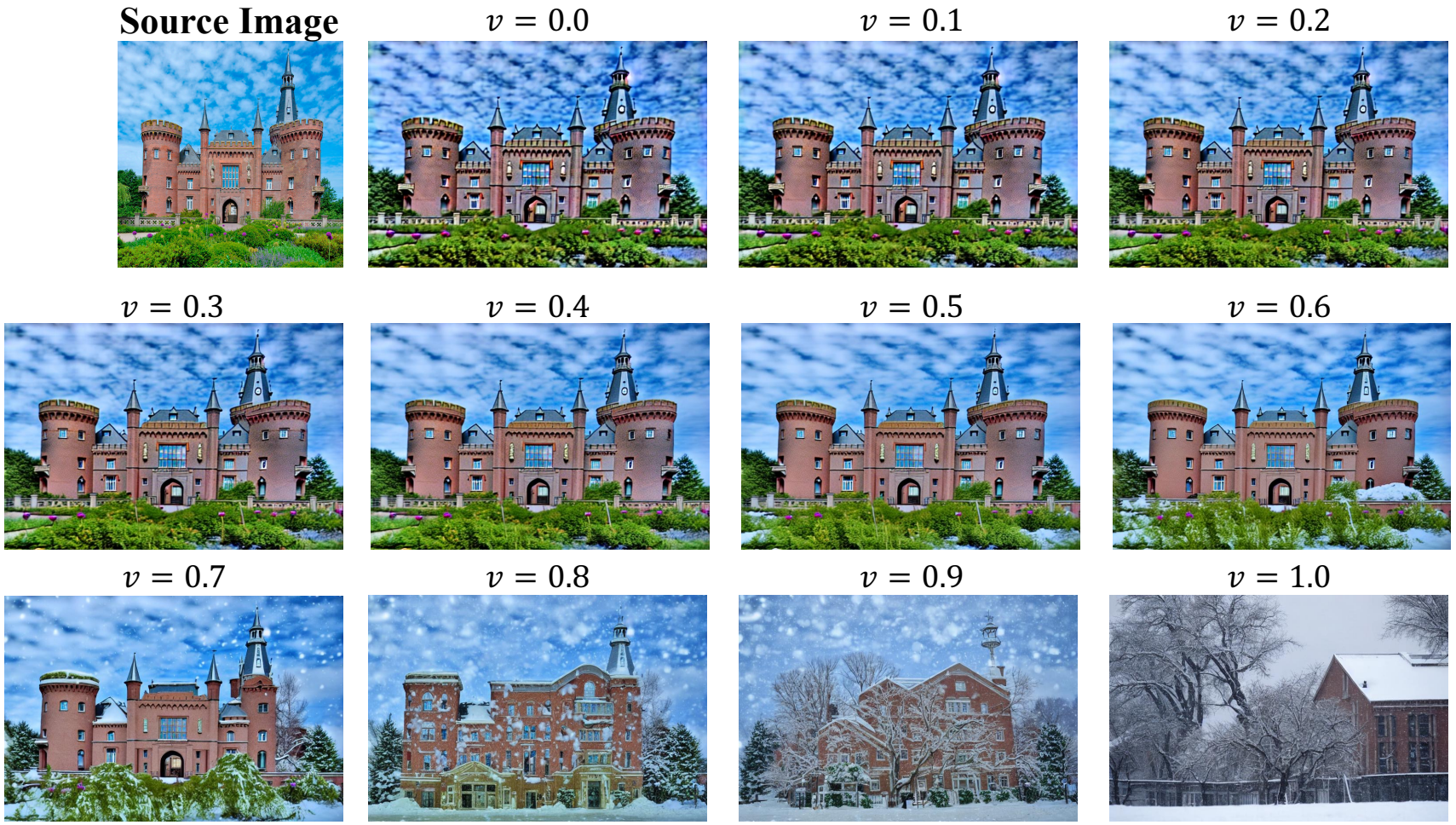


Source Image

"a sculpture of a dog in the style of Michelangelo" ($H = 768,\ W = 768$)

Analysis: Guidance weight v

- $K = 400$



Thanks!